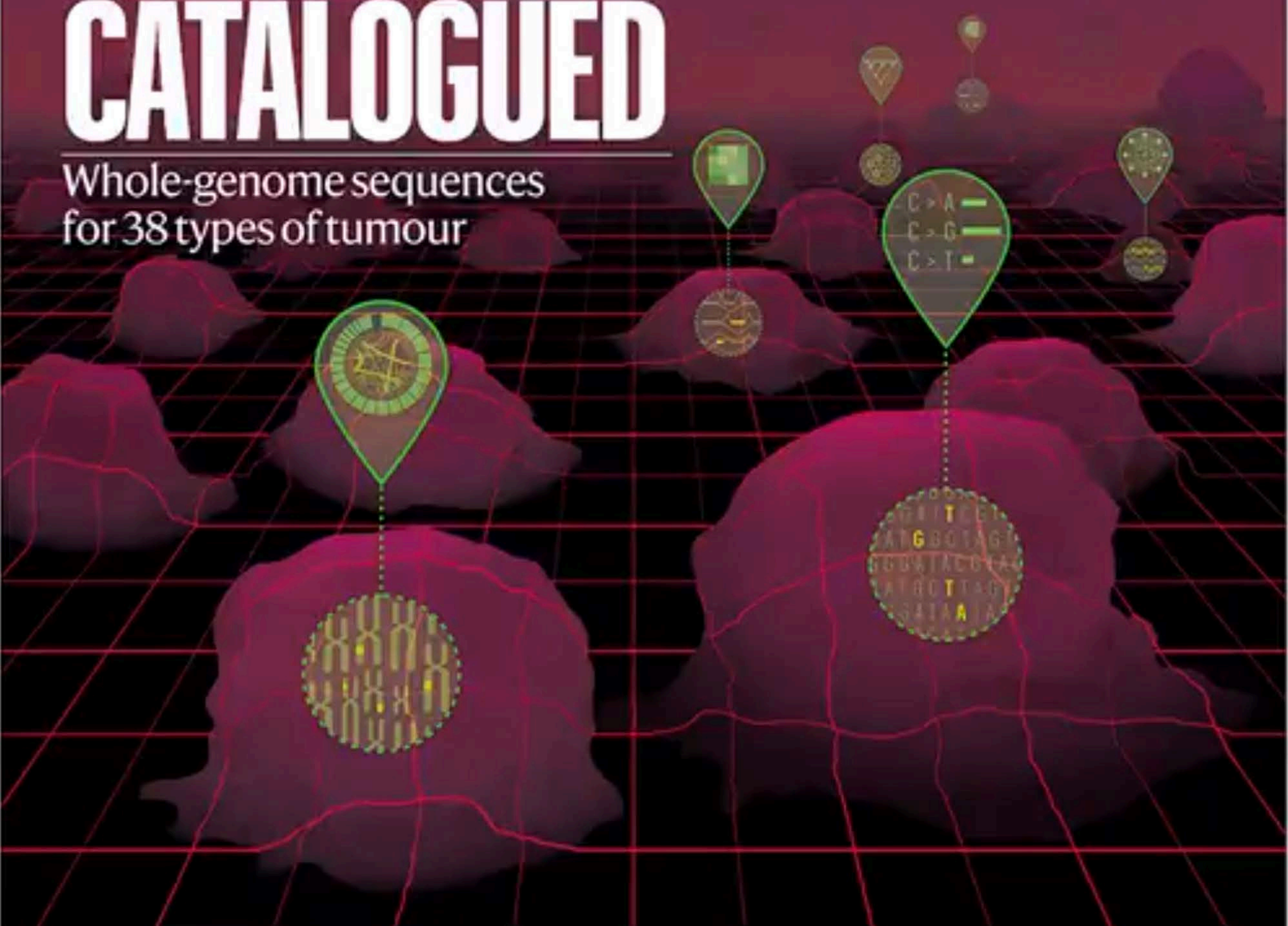


# nature

## CANCER CATALOGUED

Whole-genome sequences  
for 38 types of tumour



**Code breaker**  
RNA editing shows  
therapeutic promise  
as CRISPR alternative

**Particle alignment**  
Muon collider edges  
closer to reality with  
ionization cooling

**Going viral**  
Shock treatment  
reactivates latent HIV  
hidden in cells



## Coronavirus: keep sharing research

**Researchers must ensure that their work on this outbreak is shared rapidly and openly.**

**T**wenty thousand cases; more than 400 lives lost. The coronavirus first reported last December is now a public-health emergency of international concern. In China, cities have been sealed off, and the authorities have built an entire new hospital in Wuhan, where the outbreak started.

Along with medical workers, the country's researchers are playing a vital part. Epidemiologists are working to update estimates of case numbers; genome samples of the pathogen are being sequenced and results are being shared.

In two papers in *Nature*, teams led by researchers at the Wuhan Institute of Virology and at Fudan University, Shanghai, confirm that the virus is similar to the one that caused severe acute respiratory syndrome (SARS), and that there's evidence it originated in bats. The Wuhan team analysed viral-genome samples from a small number of patients, all of whom worked at the animal market from which the first cases were reported (P. Zhou *et al. Nature* <http://doi.org/ggj5cg>; 2020). The Fudan team sequenced a sample from one infected market worker (F. Wu *et al. Nature* <http://doi.org/dk2w>; 2020).

In the first days after the outbreak became known, we confirmed that reporting research and data will in no way affect consideration of submissions to *Nature*. *Nature* and its publisher Springer Nature have now signed a joint statement with other publishers, funders and scientific societies to ensure the rapid sharing of research data and findings relevant to the coronavirus. In the statement, we commit to working together to help ensure that:

- All peer-reviewed research publications relevant to the outbreak are made immediately open access, or freely available at least for the duration of the outbreak.
- Research findings relevant to the outbreak are shared immediately with the World Health Organization (WHO) upon journal submission, by the journal and with author knowledge.
- Research findings are made available via preprint servers before journal publication, or via platforms that make papers openly accessible before peer review, with clear statements regarding the availability of underlying data.
- Researchers share interim and final research data relating to the outbreak, together with protocols and standards used to collect the data, as rapidly and widely as possible – including with public health and research communities and the WHO.

The priority now is to stop the virus's spread and help those affected. That includes understanding how the virus is transmitted between people, ramping up

supplies of diagnostic equipment and accelerating vaccine development. Beyond this, questions are being asked about whether there were delays in sounding the alarm. Answering this honestly is necessary if we are to learn lessons for next time. It's also essential to improve regulation of animal markets, because lax oversight increases the risk of new viruses transferring from animals to humans. And funds must be released for better disease surveillance in the poorest countries – the main reason the WHO declared the virus a public-health emergency of international concern.

For researchers, the message is simple: work hard to understand and combat this infectious disease; make that work of the highest standard; and make results quickly available to the world.

## Cancer genomics gets new focus

**To realize the full potential of cancer genomics studies, tumour sequence data needs to be paired with clinical background information.**

**T**his week, *Nature* is publishing a suite of papers that sheds new light on the genetic causes of cancer. The results show how far our understanding of cancer has come – and how far we still have to go.

The Pan-Cancer Analysis of Whole Genomes Consortium brought together researchers with nearly 750 affiliations across 4 continents. Between them, they sequenced full genomes from more than 2,600 samples representing 38 different types of cancer. The work is summarized in a News & Views article on page 39.

The project is remarkable in both scope and complexity, and, partly because of this, faced challenges at every step; from acquiring samples to protecting patient privacy while putting terabytes of data into the hands of researchers.

Thanks to these efforts – and previous full-genome sequences – scientists now have an unprecedented view of the genetic changes that can contribute to cancer, and a clearer idea of where gaps in knowledge remain. Altogether, the team pinpointed 705 mutations that occurred repeatedly in the cancer genomes, suggesting that they are important for tumour growth. Of these, about 100 fell outside the protein-coding regions of the genome, but more such mutations might be uncovered with improvements in computational techniques for analysing non-coding regions. Overall, the authors found that cancer genomes contain an average of four to five mutations that drive tumour growth. In 5% of cases, however, they found no such mutations.

Cancer genomes have been sequenced for more than a decade, but now researchers and the funders who support them must tackle the next challenge. The goal has always

 **The priority now is to stop the virus's spread and help those affected."**



been to improve the lives of those affected by cancer, and the reams of data amassed by sequencing projects have helped. They are used by researchers to find new drug targets, and to generate new markers that can be used to match patients with the treatment most likely to help.

But most of the data so far have been limited in one crucial respect: clinical details of the sample donors are often missing. The first samples collected for the Cancer Genome Atlas, a sequencing project that ran from 2006 to 2018, co-funded by the US National Cancer Institute and the National Human Genome Research Institute, typically came with little more than the donor's gender, diagnosis and age at diagnosis. Rarely would there be a record of that person's family or medical history, what therapy they had received and how they had responded – all crucial information if genome sequences are to be put to work to help patients.

The next generation of cancer-genome sequencing projects is trying to change that. But gathering detailed clinical information is more difficult – and more expensive – than sequencing genomes, particularly in the many countries that lack a unified health-care system. There, accessing hospital records is complicated: different hospitals keep records differently; patients often move from one treatment centre to another; and the quality of records varies enormously. More-detailed records also mean greater risk of personal exposure if there is a privacy violation, raising the bar yet again for participant protection.

These are all pressing issues, not only in cancer research, but in health care generally. Efforts are already under way to transform health records into a format that can be more readily, but securely, accessed and studied. The American Association for Cancer Research's project GENIE, for example, has compiled 70,000 records of tumour DNA sequences, and real-world clinical data. The United Kingdom's 100,000 Genomes Project also aims to match DNA sequences with clinical information for a variety of conditions. And the International Cancer Genome Consortium, which has coordinated much of the tumour sequencing work so far, has launched a new phase, this time with a focus on clinical information.

Pooling large numbers of samples is a powerful way to find genetic changes that can drive cancer, and provides a starting point for learning how they do so. But the real return on investment will come when that information can be used to tailor therapy to individual patients. And for that to be achieved, clinical background information on study participants is essential.

When cancer-genome sequencing projects were first launched, it was hoped that they would provide a catalogue of mutations that could give rise to cancer – and reveal broad patterns on which researchers could base drug development. The core of that mission has been achieved, but many cancers have proved more complex than expected. Seemingly similar cancers can contain very different sets of mutations – no two cancers are quite the same.

As is often the case in biomedical research, the answers to a question are more complex than originally imagined. But recognizing the complexity is empowering, and harnessing it will be necessary in the search for better treatments.

“  
The answers  
are more  
complex  
than  
originally  
imagined.”

## Read all about it

***Nature* will trial the publication of peer-review reports.**

**R**esearch communities are unanimous in acknowledging the value of peer review, but there's a growing desire for more transparency in the process. As part of that, researchers want to see how publishing decisions are made, and they want greater assurance that referees and editors act with integrity and without bias.

For many journals, including *Nature*, peer review has typically been single-blind – that is, authors do not know who is reviewing their paper. At the same time, the contents of peer-review reports, and correspondence between authors, reviewers and editors, are kept confidential.

This prevents readers from seeing the often fascinating and important discussions between authors and reviewers, which are crucial in shaping and improving research and checking its integrity. Keeping these debates confidential also helps to reinforce perceptions that the research paper is the last word on a subject – when the latest finding is often simply a milestone along the scholarly journey.

Our authors have told us they want change. In a 2017 survey of *Nature* referees, 63% of respondents said publishers should experiment with alternative models, and more than half said peer review could be more transparent.

Four years ago, *Nature* invited referees to be acknowledged in papers – with the consent of both author and reviewer. Around 3,700 *Nature* referees have chosen to be publicly recognized, and around 80% of the journal's papers have at least one referee named.

Beginning this week, authors of new submissions to *Nature* will be offered the option to have anonymous referee reports published, along with their own responses and rebuttals, once a manuscript is ready for publication.

Those who agree to act as reviewers should know that their anonymous reports – and their anonymized correspondence with authors – might be published. Referees can also choose to be named, should they desire.

In making this change, *Nature* is following seven other Nature Research journals. And we're joining the pioneering efforts of *The EMBO Journal* and BMC journals – and, more recently, *Nature Communications*, which has been publishing reviewer reports since 2016.

We will report back as the trial progresses, but the experience of *Nature Communications* has been positive. In 2018, the overwhelming majority (98%) of the journal's authors who had published their reviewer reports told us they would do so again.

Published peer reviews are intended to advance scholarly discussion about a piece of research and it is important that our readers and the research community at large can benefit from such discourse. We are pleased to be playing a small part in making that happen.



been to improve the lives of those affected by cancer, and the reams of data amassed by sequencing projects have helped. They are used by researchers to find new drug targets, and to generate new markers that can be used to match patients with the treatment most likely to help.

But most of the data so far have been limited in one crucial respect: clinical details of the sample donors are often missing. The first samples collected for the Cancer Genome Atlas, a sequencing project that ran from 2006 to 2018, co-funded by the US National Cancer Institute and the National Human Genome Research Institute, typically came with little more than the donor's gender, diagnosis and age at diagnosis. Rarely would there be a record of that person's family or medical history, what therapy they had received and how they had responded – all crucial information if genome sequences are to be put to work to help patients.

The next generation of cancer-genome sequencing projects is trying to change that. But gathering detailed clinical information is more difficult – and more expensive – than sequencing genomes, particularly in the many countries that lack a unified health-care system. There, accessing hospital records is complicated: different hospitals keep records differently; patients often move from one treatment centre to another; and the quality of records varies enormously. More-detailed records also mean greater risk of personal exposure if there is a privacy violation, raising the bar yet again for participant protection.

These are all pressing issues, not only in cancer research, but in health care generally. Efforts are already under way to transform health records into a format that can be more readily, but securely, accessed and studied. The American Association for Cancer Research's project GENIE, for example, has compiled 70,000 records of tumour DNA sequences, and real-world clinical data. The United Kingdom's 100,000 Genomes Project also aims to match DNA sequences with clinical information for a variety of conditions. And the International Cancer Genome Consortium, which has coordinated much of the tumour sequencing work so far, has launched a new phase, this time with a focus on clinical information.

Pooling large numbers of samples is a powerful way to find genetic changes that can drive cancer, and provides a starting point for learning how they do so. But the real return on investment will come when that information can be used to tailor therapy to individual patients. And for that to be achieved, clinical background information on study participants is essential.

When cancer-genome sequencing projects were first launched, it was hoped that they would provide a catalogue of mutations that could give rise to cancer – and reveal broad patterns on which researchers could base drug development. The core of that mission has been achieved, but many cancers have proved more complex than expected. Seemingly similar cancers can contain very different sets of mutations – no two cancers are quite the same.

As is often the case in biomedical research, the answers to a question are more complex than originally imagined. But recognizing the complexity is empowering, and harnessing it will be necessary in the search for better treatments.

“  
The answers  
are more  
complex  
than  
originally  
imagined.”

## Read all about it

**Nature will trial the publication of peer-review reports.**

Research communities are unanimous in acknowledging the value of peer review, but there's a growing desire for more transparency in the process. As part of that, researchers want to see how publishing decisions are made, and they want greater assurance that referees and editors act with integrity and without bias.

For many journals, including *Nature*, peer review has typically been single-blind – that is, authors do not know who is reviewing their paper. At the same time, the contents of peer-review reports, and correspondence between authors, reviewers and editors, are kept confidential.

This prevents readers from seeing the often fascinating and important discussions between authors and reviewers, which are crucial in shaping and improving research and checking its integrity. Keeping these debates confidential also helps to reinforce perceptions that the research paper is the last word on a subject – when the latest finding is often simply a milestone along the scholarly journey.

Our authors have told us they want change. In a 2017 survey of *Nature* referees, 63% of respondents said publishers should experiment with alternative models, and more than half said peer review could be more transparent.

Four years ago, *Nature* invited referees to be acknowledged in papers – with the consent of both author and reviewer. Around 3,700 *Nature* referees have chosen to be publicly recognized, and around 80% of the journal's papers have at least one referee named.

Beginning this week, authors of new submissions to *Nature* will be offered the option to have anonymous referee reports published, along with their own responses and rebuttals, once a manuscript is ready for publication.

Those who agree to act as reviewers should know that their anonymous reports – and their anonymized correspondence with authors – might be published. Referees can also choose to be named, should they desire.

In making this change, *Nature* is following seven other Nature Research journals. And we're joining the pioneering efforts of *The EMBO Journal* and BMC journals – and, more recently, *Nature Communications*, which has been publishing reviewer reports since 2016.

We will report back as the trial progresses, but the experience of *Nature Communications* has been positive. In 2018, the overwhelming majority (98%) of the journal's authors who had published their reviewer reports told us they would do so again.

Published peer reviews are intended to advance scholarly discussion about a piece of research and it is important that our readers and the research community at large can benefit from such discourse. We are pleased to be playing a small part in making that happen.



# World view

## People will not trust unkind science

**A mean and aggressive research working culture threatens the public's respect for scientists and their expertise, says Gail Cardew.**

**E**arlier this month, a survey from Wellcome in London confirmed that unkindness, and worse, is pervasive in science (see [go.nature.com/2v4fn3w](https://go.nature.com/2v4fn3w)). Academic leaders expressed alarm – both for the health of young researchers and for how such pressure could erode the quality of science. I think there is more to worry about.

What hope is there for those in science to build a trusting and respectful relationship with the public when so many scientists are schooled in a culture lacking these qualities?

The need for trust and respect is particularly acute now, when people, as the British politician Michael Gove infamously put it, “have had enough of experts”. Similar arguments have come from around the world.

According to a 2019 report by public-opinion research firm Ipsos Mori, the way people behave, especially their ability to think of others' interests, influences their trustworthiness. Competence is not enough ([go.nature.com/37lydga](https://go.nature.com/37lydga)). This is backed up by a survey of people living on potentially contaminated land, which found that citizens who said they did not trust the underlying science were not questioning scientists' expertise, but whether scientists shared the public's interest ([go.nature.com/2giuvyb](https://go.nature.com/2giuvyb)).

Unkindness in science saddens me for many reasons. Obviously, I feel for the devoted researchers who began their careers expecting to revel in the joy of discovery, only to find their love of the subject squeezed out, replaced by fear and anxiety. It also saddens me because I've witnessed some of this toxic culture spill out of the laboratory, into scientists' dealings with the public.

I have spent decades examining the relationship between science, culture and society, most recently as director of science and education at the Royal Institution of Great Britain in London, heading a team that connected leading scientists with the public – in person, online, on television and in the classroom. I have long believed that scientists have a duty to discuss their work and its implications.

Conducting research responsibly includes engaging with the public. More and more researchers are now making that effort: speaking at science festivals, giving public talks and visiting schools. They often describe not just their research, but how amazing it is to be a scientist, with the opportunity to think about the many unanswered questions facing humanity.

Most speakers take an interest in their audience and give thoughtful, sensitive answers to audience questions. But some become confrontational at any remark interpreted

**A kinder research culture will build stronger, deeper support for research.”**

**Gail Cardew** is vice-president of EuroScience, a non-profit association of researchers, and honorary doctor of science at the University of Sussex, Brighton, UK. She is a professor of science, culture and society, formerly at Britain's Royal Institution. Twitter: @gailsci



By Gail Cardew

as questioning their expertise. Some dismiss questions they deem irrelevant or stupid. Some take umbrage if there are no questions, sometimes mistaking diffidence for a lack of interest. Those who do engage, but unkindly, can make matters worse. In one toe-curlingly awkward case in Europe, a speaker berated women in the audience for not asking questions. I'd bet that that audience subsequently felt less consideration both for science and for the importance of policies informed by it.

I recall how a child asked Ellen Stofan, then NASA's chief scientist, how useful Lego blocks would be to get to Mars. Stofan's warm, inspiring answer: “Everything we do at NASA, someone has to imagine first,” she said. “You have to learn to be creative, to be innovative, and that's why the arts are an important part of education.” I'm sure that response left her listeners with a higher esteem for science. And Fields Medal winner Cédric Villani once told me this of public engagement: “It reinvigorates you. It also helps you to understand what you are doing and why you are doing it.”

Wellcome's survey found that nearly four-fifths of researchers think competition has created unkind and aggressive conditions. Most (61%) have witnessed bullying or harassment, and 43% have experienced it themselves. Only 37% feel comfortable speaking up.

In a *Nature* poll following up on the survey, large majorities said that institutions, funders and lab heads should be the ones responsible for changing the culture ([go.nature.com/36j4yar](https://go.nature.com/36j4yar)). If we want to build trust in science and scientists, it is not enough to think about ‘what’ we achieve; we must think about ‘how’ we influence those around us. That's why, when I left the Royal Institution, the farewell message that meant the most commended me for having achieved amazing things in a way that was “kind and humane”.

A humane environment comes about through decisions, not luck. Make time for regular reflection on how you could have handled situations better, and have the courage to admit that swiftly to those concerned. Ask people for guidance, especially those supposedly less experienced, and definitely those less powerful, because they can often provide a fresh perspective. Give people time to learn and grow; recognize when they need help and also when they need to be left alone to make their own way, including mistakes. Above all, everyone should feel able to bring their whole selves to work, where differences in lives and backgrounds are celebrated, where unique perspectives and contributions are valued and not interpreted as criticism.

A kinder research culture will build stronger, deeper support for research, as well as higher-quality science. Maintaining public trust should not mean shouting more loudly in a noisy world. Instead, let's look at our own behaviour and ask ourselves – are we really acting in the best interests of others?



# News in brief

## CORONAVIRUS PAPERS APPEARING RAPIDLY AS RESEARCHERS RESPOND TO OUTBREAK

More than 70 research studies on the new coronavirus have been released over the past few weeks, as scientists rush to understand the pathogen and how it spreads.

The virus, known as 2019-nCoV, causes a serious respiratory illness and has so far infected more than 20,000 people and killed at least 400, according to reports as *Nature* went to press. It has also spread to multiple other countries. The infection is thought to have originated in a food market in the Chinese city of Wuhan, which has been on lockdown – with travel into and out of the city restricted – since 23 January.

The escalating outbreak has prompted a flurry of research activity on the coronavirus, which emerged in humans last December and is new to science. *Nature* searched for studies about the virus using the terms ‘coronavirus’ or ‘ncov’ on the preprint servers bioRxiv, medRxiv and ChemRxiv, as well as on Google Scholar, the discussion forum virological.org, scholarly-activity tracker

Altmetric and the websites of institutions that had published preliminary research reports on the subject. As of 4 February, at least 77 English-language papers on the coronavirus have been published (see ‘Coronavirus research’).

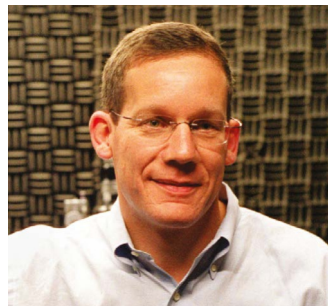
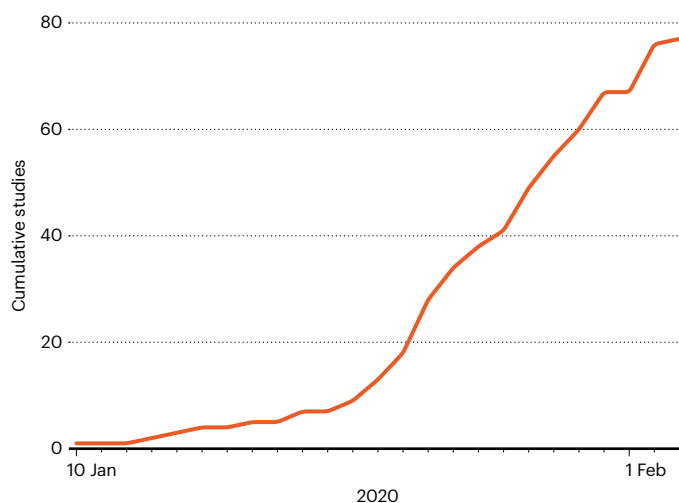
More than half of the studies are on preprint servers, and a handful have appeared in peer-reviewed journals, including *The Lancet* and the *Journal of Medical Virology*. The search did not include Chinese-language journals.

Several of the papers contain estimates of how rapidly the virus spreads, or the length of its incubation period – how long after being infected with the virus people start to experience symptoms.

Other studies focus on the virus’s structure or genetic make-up – information that could be used to identify drug targets or develop a vaccine. Researchers have also published genomic data on the virus on online platforms such as GISAID or GenBank, but *Nature*’s analysis did not count these data uploads.

## CORONAVIRUS RESEARCH

Dozens of studies about the virus have been published since the outbreak began.



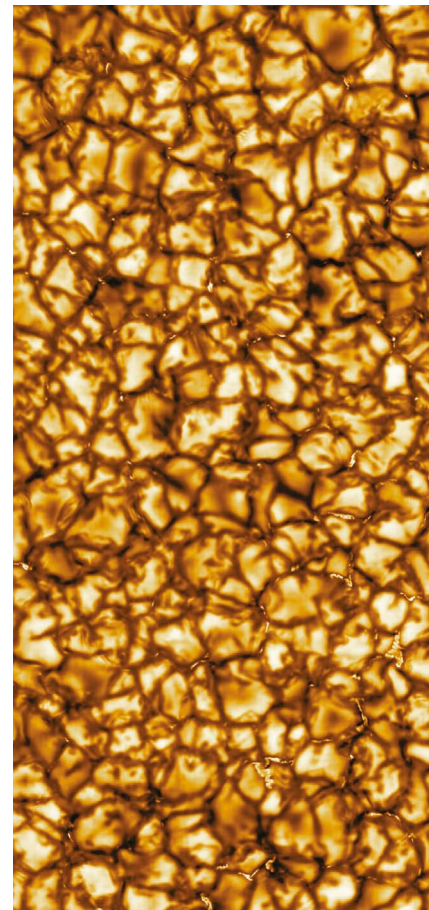
## HARVARD CHEMISTRY CHIEF'S ARREST STUNS SCIENTISTS

Researchers have reacted with shock to the arrest of top nanoscientist Charles Lieber, who has been charged with lying to the US government about receiving funding from China.

Lieber, who leads the chemistry department at Harvard University in Cambridge, Massachusetts, was arrested on 28 January and released on bail two days later.

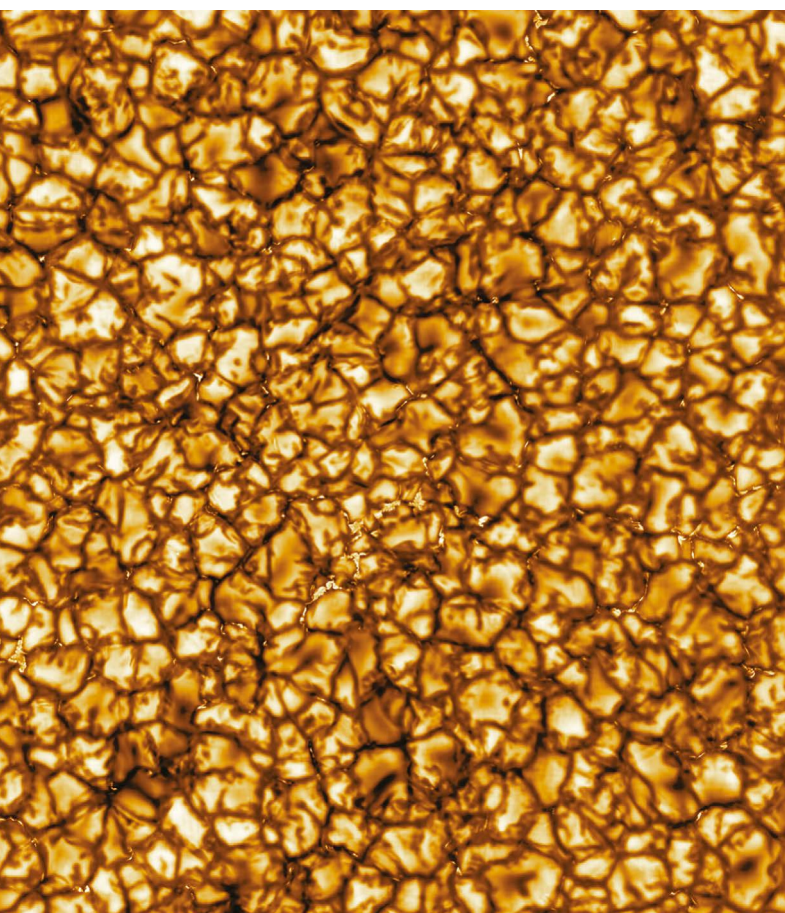
Colleagues of Lieber said they were shocked by his detainment. “I have 100% trust and confidence in him. I think there must be some misunderstanding,” says Xiaocheng Jiang, a former student of Lieber who is now at Tufts University in Medford, Massachusetts.

The federal charges focus on Lieber’s alleged involvement in China’s Thousand Talents Plan, a programme designed to recruit academics to the country. They come as US authorities increasingly scrutinize universities’ foreign ties, amid fears that countries might be stealing or influencing US research. The FBI alleges that Lieber received hundreds of thousands of dollars from a Chinese university and agreed to lead a lab there – but that he denied his involvement when asked by US government agencies. Lieber’s legal team did not respond to *Nature*’s requests for comment.



## The highest-resolution image of the Sun ever taken





The world's most powerful solar telescope has opened its eyes. The US\$344-million Daniel K. Inouye Solar Telescope, which has been two decades in the making, is scrutinizing the Sun in extraordinary detail from atop Haleakala mountain in Hawaii.

Images released on 29 January show patterns of superheated gas churning on the Sun's surface. Bright 'cells' represent the plasma rising from deeper in the star, and the darker borders between the cells indicate where plasma is cooling and sinking.

The 4-metre Inouye telescope eclipses what had been the world's largest solar telescope, a 1.6-metre facility at Big Bear Solar Observatory in southern California. Scientists say that the dramatic upgrade will transform solar physics for decades. The Inouye Solar Telescope will make the most precise measurements of the Sun's magnetic field so far, including the first-ever magnetic measurements in the Sun's atmosphere, or corona.

"It's going to be such a revolution," says Momchil Molnar, a solar physicist at the University of Colorado Boulder.

## PROMISING HIV VACCINE FAILS IN LARGE TRIAL

The quest to develop a vaccine against HIV has been dealt a setback. Researchers running a trial of a once-promising vaccine in South Africa have stopped administering immunizations after an analysis showed that the vaccine was not effective. The study's sponsor, the US National Institute of Allergy and Infectious Diseases, announced the trial's cancellation on 3 February.

The trial, called HVTN 702, enrolled 5,407 people who did not have HIV, and they received either the vaccine or a placebo injection. The vaccine that participants received was similar to one that, in a previous trial in Thailand, had reduced infections by about 30% compared with the trial's placebo group. That marked the first-ever success for an HIV vaccine in a large trial, albeit a modest one.

But an independent board that was monitoring interim data from the South Africa trial determined that, after most of the volunteers had been in the study for 18 months or more, the vaccine was not protecting participants from HIV infection. Among the 2,694 people who received the immunization, 129 contracted HIV; 123 of the 2,689 participants who received the placebo tested positive for HIV. Researchers will continue to follow the volunteers and try to determine why the vaccine failed.



## INDIA BETS BIG ON QUANTUM TECHNOLOGIES

Quantum technology has been given a massive boost in India's latest budget, receiving 80 billion rupees (US\$1.12 billion) over 5 years as part of a new national quantum mission.

The move places India alongside the United States and Europe, which in the past few years have each pledged more than US\$1 billion to research in the field. Russia also announced an initiative worth hundreds of millions of dollars late last year.

India's investment – announced on 1 February by finance minister Nirmala Sitharaman (pictured) – will be administered by the ministry of science and technology, and is a considerable increase on past commitments. A national quantum-technology research programme announced in 2018 received US\$27.9 million over 5 years.

Ashutosh Sharma, secretary of the ministry's department of science and technology, says India's quantum research is solid on the theoretical side, but needs infrastructure and experimental facilities. The mission will develop quantum technologies for communications, computing, materials development and cryptography, and will coordinate the work of scientists, industry leaders and government departments, he says.

# News in focus



Workers disinfect areas of Qingdao, China, as part of virus-control measures.

## WHAT'S NEXT FOR THE CORONAVIRUS OUTBREAK CAUSING GLOBAL ALARM

Many aspects of the new virus remain unknown, but scientists are weighing up the best- and worst-case scenarios for the future.

By Dyani Lewis

**A**s health authorities around the world race to halt the spread of a new virus that emerged in the Chinese city of Wuhan in December, scientists are considering what is likely to happen next, and how bad the outbreak might get. As *Nature* went to press, more than 20,000 people had contracted the coronavirus, which causes respiratory illness; the death toll was 426, and rising daily. Crucial details about the virus and how it spreads are still unknown, but scientists are considering best- and worst-case scenarios on the basis of previous epidemics and what they already know about this one.

### How many people will be infected?

Chinese authorities have locked down cities at the centre of the epidemic, and researchers were quick to share data on the virus with the World Health Organization (WHO) and other scientists. But the case numbers have been rising. According to one prediction, the virus could infect about 46,000 of the 30 million people living in and around Wuhan (see [go.nature.com/31m9fzz](https://go.nature.com/31m9fzz)). In the best case, fewer people will be infected because the effects of the control measures will start kicking in, says Ben Cowling, an epidemiologist at the University of Hong Kong. But it's too early to tell whether efforts to quarantine people, and the widespread use of face masks, are

working. The incubation period for the virus – up to 14 days – is longer than most control measures have been in place, he says.

In a worst-case scenario, some 190,000 people could be infected in Wuhan, according to another model (J. M. Read *et al.* Preprint at medRxiv [http://doi.org/10.1101/2020.02.05.20023222](https://doi.org/10.1101/2020.02.05.20023222); 2020). Scientists are particularly concerned about fresh outbreaks emerging outside China. The virus has already spread in small, localized clusters in Vietnam, Japan, Germany and the United States, but authorities have been quick to isolate the people affected. More than 150 cases had been recorded outside China as of 3 February, including one death in the Philippines.



## News in focus

### Is the virus here to stay?

When a virus circulates continuously in a community, it is said to be endemic. The viruses that cause chicken pox and influenza are endemic in many countries, for example. If efforts to contain the coronavirus fail, and if the virus can be passed on by people who are infected but don't have symptoms, it will be more difficult to control its spread, making it more likely that it will become endemic, too.

As with influenza, this could mean that deaths occur every year as the virus circulates, until a vaccine is developed. But if control measures are effective, and transmission slows so that each infected person infects no more than one other person, the current outbreak could simply peter out, says Cowling.

There have been several cases of infected people displaying no symptoms, but it's still unclear whether such asymptomatic or mild cases are common, and whether or to what extent they are infectious. Asymptomatic cases set the new virus apart from the related coronavirus that causes severe acute respiratory syndrome (SARS). There was a global outbreak of the SARS virus in 2002–03, but it usually spread only once people were ill enough to need hospital care, and was eventually contained. There is no evidence that the virus is still circulating in humans, says Ian Mackay, a virologist at the University of Queensland in Brisbane, Australia.

### Is the new virus likely to change?

Currently, it has caused severe illness, and death, mainly in older people, particularly those with pre-existing health conditions. Some researchers are worried that the pathogen could mutate so it can spread more efficiently, or become more likely to cause disease in young people.

But Kristian Andersen, an infectious-disease researcher at Scripps Research in La Jolla, California, says this is unlikely. Viruses constantly mutate, he says, but those mutations don't typically make the virus more virulent or cause more serious disease: most mutations are detrimental to the virus or have no effect. A 2018 study of the SARS virus in primate cells found that a mutation it sustained during the 2003 outbreak probably reduced its virulence (D. Muth *et al. Sci. Rep.* 8, 15177; 2018).

Researchers have shared dozens of genetic sequences from strains of the new coronavirus, and a steady supply of those sequences will reveal genetic changes as the outbreak progresses, says Mackay.

### How many people will it kill?

The fatality rate for a virus – the proportion of infected people who die – is difficult to calculate in the middle of an outbreak because records are constantly being updated. With about 400 deaths so far out of more than 20,000 infections, the new coronavirus has a

death rate of 2–3%. This is significantly lower than SARS, which killed around 10% of infected people. And the known death rate for the new coronavirus is likely to decrease as mild and asymptomatic cases are identified, says virologist Mark Harris at the University of Leeds, UK.

The death toll will also depend on how China's health systems cope with the number of cases. Putting people on drips and ventilators can ensure that they get enough fluids and oxygen while their immune systems fight the virus. If the virus spreads to regions with few resources, such as parts of Africa, health systems could struggle, says Sanjaya Senanayake, an infectious-disease specialist at the Australian National University in Canberra.

On 30 January, the WHO declared the outbreak a “public-health emergency of international concern”. WHO director-general Tedros Adhanom Ghebreyesus said his main

concern was that the outbreak could spread to countries with fragile health systems.

There are currently no effective drugs against the virus, but two HIV drugs are being tested as a treatment, and several research groups are working on a vaccine.

The current death rate of 2–3% – although not as high as that of SARS – is still high for an infectious disease, says Adam Kamradt-Scott, a global health-security specialist at the University of Sydney, Australia. The 1918 influenza outbreak infected around half a billion people, one-third of the world's population at the time, and killed more than 2.5% of those infected; some have estimated that as many as 50 million people died. The new coronavirus probably won't trigger such an apocalyptic scenario, because it isn't typically infecting or killing young, healthy people, says Kamradt-Scott.

## CORONAVIRUS: LABS WORLDWIDE SCRAMBLE TO ANALYSE SAMPLES

Scientists need the pathogen to develop tests, drugs and vaccines.

By Ewen Callaway

**W**ith no sign that an outbreak of a new coronavirus is abating, virologists worldwide are itching to get physical samples of the virus. They are planning to test drugs and vaccines, develop animal models of the infection and investigate the virus's biology.

“The moment we heard about this outbreak, we started to put our feelers out to get access to these isolates,” says Vincent Munster, a virologist at the US National Institute of Allergy and Infectious Diseases centre in Hamilton, Montana. His lab is expecting to receive a sample soon from the US Centers for Disease Control and Prevention in Atlanta, Georgia.

The first lab to isolate and study the virus, known as 2019-nCoV, was at the Wuhan Institute of Virology – in the city where the outbreak started. A team led by virologist Zheng-Li Shi isolated the virus from a woman who developed symptoms in December. Shi's team found that the virus can kill cultured human cells and that it enters them through the same molecular receptor as the coronavirus that causes SARS (severe acute respiratory syndrome; P. Zhou *et al. Preprint at bioRxiv* <http://doi.org/ggjs7d>; 2020).

An Australian lab said on 28 January that it

had obtained virus samples from an infected person who had returned from China, and it was preparing to share them with other scientists. Labs in France, Germany and Hong Kong are also isolating and preparing to share virus samples taken from local patients, says Bart Haagmans, a virologist at Erasmus Medical Center in Rotterdam, the Netherlands. Haagmans hopes to receive viral material from one of these labs in the coming days.

Several labs have sequenced the virus, but scientists say that the results are no substitute for live samples, which are needed to test drugs and vaccines, and to study the virus in depth. Munster says that his priority will be to identify animals that experience the infection in a similar way to humans. These will be useful for developing vaccines and drugs. The team first plans to look at a mouse genetically engineered to contain a human version of the receptor that the coronavirus uses to infect cells. Future work could involve exposing mice and non-human primates to the virus and testing whether vaccines can prevent infection.

Munster's lab is also eager to start gauging how long the virus can survive in the air or in saliva droplets. This could help epidemiologists to understand whether the virus can be transmitted through the air, or only through close contact.

# SOCIAL SCIENTISTS BATTLE BOTS TO GLEAN INSIGHTS ONLINE

Automated production of social-media posts can confound research studies.

By Heidi Ledford

**S**ocial-media bots that pump out computer-generated content have been accused of swaying elections and damaging public health by spreading misinformation. Now, some social scientists have a fresh accusation: bots meddle with research studies that mine popular sites such as Twitter, Reddit and Instagram for information on human health and behaviour.

Data from these sites can help scientists to understand how natural disasters affect mental health, why young people have flocked to e-cigarettes in the United States and how people join together in complex social networks. But such work relies on distinguishing the real voices from the automated ones.

"Bots are designed to behave online like people," says Jon-Patrick Allem, a social scientist at the University of Southern California in Los Angeles. "If a researcher is interested in describing public attitudes, you have to be sure that the data you're collecting on social media is actually from people."

Computer scientist Sune Lehmann designed his first bots in 2013, as a social-network experiment for a class that he was teaching at the

Technical University of Denmark in Kongens Lyngby. Back then, he says, Twitter bots were simple, obscure and mainly meant to increase the number of followers for specific accounts. Lehmann wanted to show his students how such bots could manipulate social systems, so together they designed simple bots that impersonated fans of the singer Justin Bieber.

The 'Bieber Bots' quickly attracted thousands of followers. But social-media bots have continued to evolve, becoming more complex and harder to detect. They surged into the spotlight after the 2016 US presidential election – amid accusations that bots had been deployed on social media in an attempt to sway the vote. "All of a sudden, it became something of interest to people," Allem says.

Since then, Allem has shown that tweets generated by bots are twice as likely as their real counterparts to attest that e-cigarettes help people to give up smoking<sup>1</sup>. Bots are also more likely to tout the unproven health benefits of cannabis<sup>2</sup>. These studies rely on algorithms that estimate the likelihood that a Twitter account is automated. But despite bot-detecting tools with names like BotSlayer, Allem says that many social-science and public-health researchers still fail to filter out

probable automated content from their data.

That omission can pollute a data set, says Amelia Jamison, who studies health disparities at the University of Maryland in College Park and has mined social media for posts that oppose vaccination. "You might be artificially giving the bots a voice by treating them as if they are really part of the discussion, when they are actually just amplifying something that may not be voiced by the community," she says.

One problem that the field must grapple with is how to define a bot, says Katrin Weller, an information scientist at the Leibniz Institute for the Social Sciences in Cologne, Germany. Not all bots are dispensing misinformation: some provide data from weather stations, or general news updates. Some researchers define Twitter bots as those accounts that send out more than a certain number of messages each day – a loose definition that could rope in prolific human tweeters.

Other definitions are more complex, but bot detectors are locked in an arms race with bot developers. First-generation social-media bots were relatively simple programs that retweeted others' posts at regular intervals. Now, advances in machine learning have enabled the creation of more sophisticated bots that post original content. Some post at random intervals and mimic human patterns, such as not tweeting when a person would probably be asleep. Some developers will mix in human-generated content with automated content to better camouflage their bots.

"Once you know more about the bots and how to detect them, then this knowledge is also available for the bot creators," says Oliver Grübner, who studies quantitative health geography at the University of Zurich in Switzerland. "It's a really tricky field."

Like Lehman, some social scientists are creating their own bots to conduct social experiments. Kevin Munger, a political scientist at Pennsylvania State University in University Park, and his colleagues built bots that chided Twitter users who used racist language. One set of bots had profile pictures of white men; the other had profile pictures of black men. Munger found that Twitter users were more likely to tone down their racist rhetoric after being called out by bots with a white male profile picture<sup>3</sup>.

After his Bieber Bot success, Lehmann designed more sophisticated bots to study how behaviours spread from one group to another. But bots now have such a bad reputation that he is leaning towards abandoning the approach, for fear of a public backlash. "I kind of thought: 'I'll find another quiet corner and do my research without courting controversy,'" he says.

1. Allem, J.-P. et al. *JMIR Public Health Surveill.* **3**, e98 (2017).

2. Allem, J.-P. et al. *Am. J. Public Health* <https://doi.org/10.2105/AJPH.2019.305461> (2019).

3. Munger, K. *Political Behav.* **39**, 629–649 (2017).



Separating real online voices from automated ones can be a problem for researchers.



## How quickly can Iran make a nuclear bomb?

**With an international deal in jeopardy, Iran is not racing to build a nuclear weapon — but its capabilities are growing again.**

Iran has accumulated 1,200 kilograms of enriched uranium — more than doubling the stockpile it had just three months ago, according to a statement from a senior official at the Atomic Energy Organization of Iran on 25 January.

That's enough to build one atomic bomb, if the uranium is further refined to make it weapons-grade — a process that could take just two to three months, says David Albright, a nuclear-policy specialist at the Institute for Science and International Security in Washington DC. Building actual weapons would take much longer, he adds.

If confirmed, the rate of expansion of Iran's uranium stockpile "shifts things dramatically", Albright adds. But he and others say that there is no evidence that Iran is rushing to build a bomb — for now.

Tensions between Iran and the United States have been escalating. On 3 January, a US drone strike killed Qasem Soleimani, the architect of Iran's military involvement in the Middle East. In response, Iran shot missiles at US bases in Iraq.

The Joint Comprehensive Plan of Action (JCPOA), the 2015 deal between Iran and six global powers that limited its nuclear capabilities in exchange for the lifting of economic sanctions, is now in serious jeopardy. US President Donald Trump pulled out of the deal in May 2018, and Iran announced in May last year that it would resume uranium enrichment.

Nature talked to nuclear experts to find out how soon Iran could build a bomb, and whether this is likely to happen.

### Has Iran tried to build nuclear weapons before?

Building nuclear weapons is expensive and requires technical expertise in enriching uranium. The fissionable isotope uranium-235, which makes up less than 1% of natural uranium, must be separated from uranium-238, which is by far the more common isotope. Iran has had an active nuclear programme for decades. The country has always maintained that this was purely for peaceful purposes,

such as producing isotopes for medical use. But in the early 2000s, Iran seemed to have a programme to build at least five uranium fission bombs, according to US intelligence assessments and international observers.

Reports in the mid-2000s by the United Nations International Atomic Energy Agency (IAEA) suggested that Iran might have been actively working to build a nuclear arsenal. That would be a violation of the 1968 Treaty on the Non-Proliferation of Nuclear Weapons (NPT), which Iran has signed. In 2003, bowing to international pressure, the country agreed to cut down its nuclear activities drastically — but not completely.

### Did the 2015 deal reduce Iran's nuclear capabilities?

By 2015, the country had stockpiled 11 tonnes of uranium hexafluoride enriched to as much as 20%  $^{235}\text{U}$  — weapons-grade uranium must be enriched to 90%. Uranium hexafluoride is processed in for enrichment in gas form, in high-speed centrifuges, and in 2015 Iran had more than 10,000 of these centrifuges. When the JCPOA was signed in July that year, experts estimated that the country was months — perhaps weeks — away from producing weapons-grade uranium.

But the JCPOA forced Iran to ship most

of its stockpile abroad, and to mothball the majority of its centrifuges. The aim was partly to stretch the time Iran needed to stockpile enough fissile material for a bomb — known as 'breakout time' — to at least a year. The deal also subjected Iran to a stringent regime of IAEA inspections.

### What was the impact of the US withdrawal from the nuclear deal?

Seyed Hossein Mousavian, who was a

## "Tensions between Iran and the United States have been escalating."

spokesperson for Iran's nuclear negotiating team in 2003, says that Iranians feel cheated. The perception in the country is that "you cannot negotiate with or trust the US", says Mousavian, now a nuclear-policy specialist at Princeton University in New Jersey.

### Does Iran now have enough enriched uranium to build nuclear bombs?

Last November, the IAEA found that Iran had accumulated around 550 kilograms of uranium hexafluoride that was



Iran's Bushehr nuclear power plant has been in operation for almost ten years.

MAJID ASGARIPOUR/AP/SHUTTERSTOCK

“moderately enriched” to less than 4.5%  $^{235}\text{U}$ . It is unclear what material the Iranian official was referring to in his 25 January claim, but it is presumed to be 1,200 kilograms of moderately enriched uranium hexafluoride. If further enriched, this could yield more than 30 kilograms of weapons-grade uranium, enough to build one fission bomb.

#### How quickly could Iran make a bomb once it has enough weapons-grade uranium?

Possessing fissile material is not enough: a country also has to master the design and manufacture of a bomb. In particular, uranium hexafluoride must be converted to uranium metal, which is not straightforward, says Richard Johnson, a proliferation specialist at the Nuclear Threat Initiative, a policy research centre in Washington DC. According to Albright, some intelligence agencies estimate that it could take Iran about two years to make its first two bombs if it wanted to do this.

#### If the nuclear deal is scrapped, will Iran be legally entitled to go nuclear?

No. Because Iran has signed the NPT, it is committed to using nuclear technology exclusively for peaceful purposes. Members of the NPT must allow the IAEA to verify their compliance, or face consequences. But Iran could withdraw from the NPT, as North Korea did in 2003, as it was becoming a nuclear power. Iranian foreign minister Mohammad Javad Zarif said on 20 January that the country is prepared to withdraw if its continued enrichment programme is reported to the UN Security Council.

#### So is Iran actively working towards a nuclear bomb?

“All the signs are that they are not,” says Zia Mian, a physicist and nuclear-policy expert at Princeton University. The country has complied with the rigorous IAEA inspection regime set out in the JCPOA. This means that a nuclear-weapons programme is “either hidden so well that no one has been able to find it so far, or that there is no such crash programme”, he says. Albright agrees, saying that Iran could be stockpiling enriched uranium to increase its leverage in future negotiations. “You don’t see some of the indicators that would imply a well worked-out decision” to actually build bombs, he says. But expanding stockpiles of enriched uranium brings the country closer to being able to make a nuclear bomb — if it wishes.

By Davide Castelvecchi



CHRISTOPHER FURLONG/GETTY

The Xhosa people have greater genetic diversity than do people of non-African descent.

## AFRICAN SCHIZOPHRENIA STUDY IDENTIFIES DAMAGING MUTATIONS

Genetic studies of mental illness have largely been conducted in people with European ancestry.

By Alison Abbott

**T**he first genomic analysis of schizophrenia in an African population has identified multiple rare mutations that occur more frequently in people with the condition.

The mutations are mainly in genes that are important for brain development and the brain’s synapses, structures that coordinate communication between neurons. The findings, published on 31 January (S. Gulsuner *et al. Science* **367**, 569–573; 2020), match those of other schizophrenia studies — but nearly all previous research has been conducted in European or Asian populations.

This research is important because Africa has represented a big gap in the populations that geneticists have studied, says psychiatric geneticist Andreas Meyer-Lindenberg, director of the Central Institute of Mental Health in Mannheim, Germany. He says that the work lends support to current hypotheses about the biological origins of schizophrenia, which can cause hallucinations and delusions. Researchers think that each mutation contributes a small amount to the overall risk of developing the condition, and that disruption to synapses could be crucial to the disease’s development.

Geneticists have long been criticized for failing to sample diverse populations for genomic

studies, which have largely neglected African people. “This urgently needs more attention,” says Ambroise Wonkam, a human geneticist at the University of Cape Town, South Africa.

This bias means that diagnostic tests and treatments developed on the basis of these studies might not work in certain populations. But studies in diverse populations also allow researchers to build up a fuller picture of diseases. In particular, African people as a group have genomes that are more diverse than those of other populations because the vast majority of human evolution took place in Africa.

The study enrolled around 900 people with schizophrenia and a similar number of controls. All identified as Xhosa, members of an ethnic group who live mainly in South Africa.

The researchers sequenced the participants’ genomes and searched for mutations that damage genes. Such mutations were much more prevalent in people with schizophrenia than in the control individuals, and were concentrated in genes that are highly expressed in the brain or involved in synapse function.

The results echo those of a large Swedish study (G. Genovese *et al. Nature Neurosci.* **19**, 1433–1441; 2016) that used the same methods, but the density of mutations in affected genes were generally larger in the Xhosa participants. The authors say that this reflects the greater genetic variation among Xhosa people.

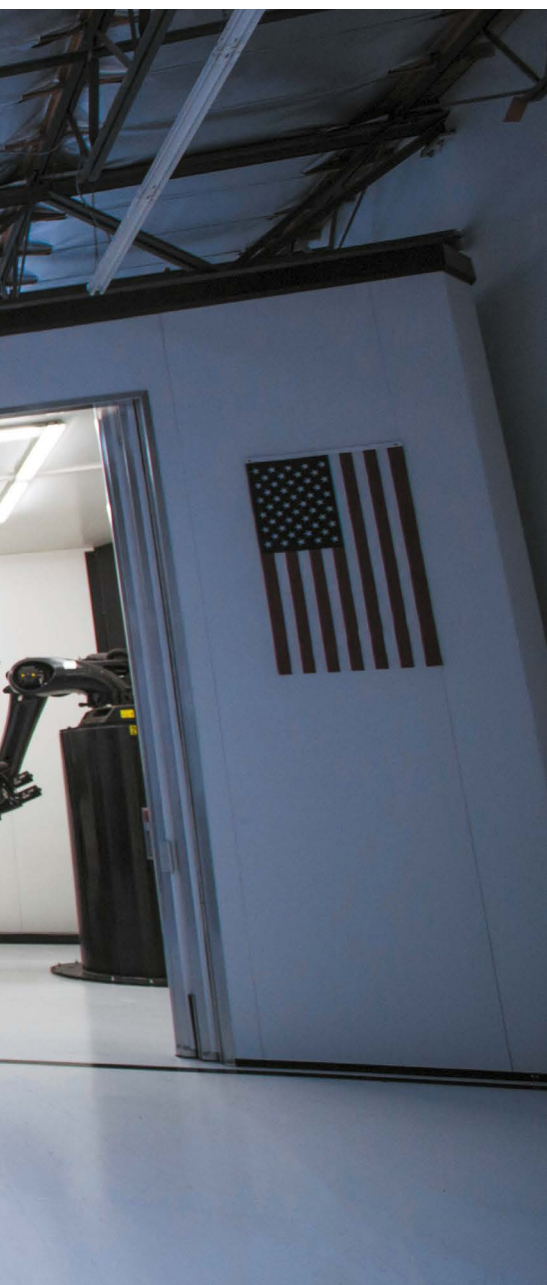




# THE NEW 3D PRINTING

Researchers are developing techniques to print faster, bigger and weirder.  
**By Mark Zastrow**

**A**s a metal platform rises from a vat of liquid resin, it pulls an intricate white shape from the liquid – like a waxy creature emerging from a lagoon. This machine is the world's fastest resin-based 3D printer and it can create a plastic structure as large as a person in a few hours, says Chad Mirkin, a chemist at Northwestern University in Evanston, Illinois. The machine, which Mirkin and his colleagues reported last October<sup>1</sup>, is one of a slew of research advances in 3D printing that are broadening the prospects of a technology once viewed as useful mainly for making small, low-quality prototype parts. Not only is 3D printing becoming faster and producing larger products, but scientists are coming up with innovative ways to print and are creating stronger materials, sometimes mixing multiple materials in the same product.



**A metal printer at start-up firm Relativity Space, which aims to test a mostly 3D-printed rocket this year.**

Sportswear firms, aviation and aerospace manufacturers and medical-device companies are eager to take advantage. “You’re not going to be sitting in your home, printing out exactly what you want to repair your car any time soon, but major manufacturing companies are really adopting this technology,” says Jennifer Lewis, a materials scientist at Harvard University in Cambridge, Massachusetts.

The latest techniques could be lucrative for researchers, many of whom – Lewis and Mirkin among them – are already commercializing their work. They’re also fundamentally exciting, says Iain Todd, a metallurgist at the University of Sheffield, UK. “We can get

performance out of these materials that we didn’t think we could get. That’s what’s really exciting to a materials scientist. This is getting people used to the new weird.”

### From trinkets to products

The 3D printing technique is also referred to as ‘additive manufacturing’, because instead of chopping or milling a shape out of a larger block, or casting molten material in a mould, it involves building objects from the bottom up. Its advantages include less waste and an ability to print custom designs, such as intricate lattice structures, that are otherwise hard to create. Low-cost hobbyist machines print by squeezing out thin plastic filaments from heated nozzles, building up a structure layer by layer – a method known as fused deposition modelling (FDM). But the term 3D printing encompasses a much wider range of techniques. One of the oldest uses an ultraviolet laser to scan across and solidify (or ‘cure’) light-sensitive resin, layer by layer. That concept was described as far back as 1984, in a patent filed by Charles Hull<sup>2</sup>, the founder of a company called 3D Systems in Rock Hill, South Carolina.

The latest techniques – including Mirkin’s – still use light-sensitive resin, but are faster and larger-scale, following improvements reported in 2015 by a team led by Joseph DeSimone, a chemist and materials scientist at the University of North Carolina at Chapel Hill<sup>3</sup>. Early printers were slow, small-scale and prone to producing layered, imperfect and weak structures. These found a niche in rapid prototyping, making plastic model parts as mock-ups for later production by conventional methods. As an area of research, this kind of printing wasn’t thrilling, says Timothy Scott, a polymer scientist at Monash University in Melbourne, Australia: “Basically making trinkets and knick-knacks. For a polymer chemist, it was pretty dull.”

Then DeSimone unveiled a way to print light-sensitive resin up to 100 times faster than conventional printers<sup>3</sup>. It uses a stage submerged in a vat of resin. A digital projector shines a pre-programmed image up at the stage through a transparent window in the floor of the vat. The light cures an entire resin layer at once. DeSimone’s advance was to make the window permeable to oxygen. This kills the curing reaction and creates a thin buffer layer, or ‘dead zone’, just above the window’s surface so that the resin doesn’t stick to the bottom of the vat each time a layer is printed. The stage rises continually, pulling the completed part up through the liquid as new layers are added at the bottom.

Other labs were working on similar concepts at the time, says Lewis. But perhaps most impressive about DeSimone’s resins was that they could undergo a second reaction in a post-print heat treatment to strengthen the

finished product. “It opens up a much broader array of materials,” says Lewis.

Many research groups and firms have since built on the work. Mirkin’s printer pumps a layer of clear oil across the bottom of the vat to inhibit the polymer’s reactions. This also acts as a coolant, removing heat that can deform a printed part – and it means that the equipment is not limited to printing with resins that are inhibited by oxygen. He says the printer produces material ten times faster than DeSimone’s. And last January, Scott and his colleague Mark Burns at the University of Michigan in Ann Arbor reported a printer that inhibits the reactions by mixing into the resin a chemical that can be activated by a second lamp emitting a different wavelength of light<sup>4</sup>. By varying the ratio of the strength of the two light sources, the researchers can control the thickness of the photo-inhibited zone, allowing the creation of more complicated patterns, such as surfaces embossed with seals or logos.

Inventions in 3D printing often have rapid commercial potential: some researchers start forming companies before they publish their

**“We can get performance out of these materials that we didn’t think we could get.”**

advances. On the same day DeSimone’s paper was published, for instance, he showcased it at a TED talk in Vancouver, Canada, and officially launched his start-up firm Carbon 3D in Redwood City, California, although he had quietly registered the company two years earlier. The firm is now one of the biggest start-ups in 3D printing; it has already raised US\$680 million in publicly disclosed funding rounds, and is reportedly valued at \$2.4 billion. It has high-profile contracts with Adidas to make rubber-like midsoles for athletic shoes, and with sports-gear firm Riddell to manufacture customized helmet padding for American-football players.

Mirkin and his colleagues James Hedrick and David Walker have also launched a start-up, Azul 3D in Evanston, Illinois, to commercialize their technique, which they have dubbed HARP (high-area rapid printing). And Scott and Burns are preparing a commercial prototype printer with their Ann Arbor-based start-up DiploDocal, a name derived from the Greek for ‘double beam’.

New resin-printing techniques are still emerging. One begins with a small spinning glass holding liquid resin. As the glass rotates, a projector shines a loop of video onto it that corresponds to 2D slices of the desired object. Within seconds, the final object solidifies inside the liquid resin – no layers necessary<sup>5</sup>. The method is inspired by X-rays and computed-tomography scans, which image



## Feature

a cross-section of a solid object. This is the inverse: back-projecting cross-sections to form a 3D object.

Even in this fast-moving field, the technique turned heads for what Lewis calls “the gee-whiz factor”. It has significant limitations: the resin used must be transparent, and the printed object must be small enough for light to pass through it to cure it. But it also has a potential advantage: it can handle highly viscous resins, which other resin-based printers struggle to suck through the narrow dead zone. That means it could make stronger materials and more accurate prints.

The approach has garnered substantial interest from industry, says Christopher Spadaccini, a materials and manufacturing engineer at Lawrence Livermore National Laboratory (LLNL) in California. Spadaccini was a member of the team that published the idea last January<sup>5</sup>, although a group at the Swiss Federal Institute of Technology in Lausanne (EPFL) independently developed the same concept and reported it in a preprint a few months beforehand<sup>6</sup>. Spadaccini thinks the technology has tremendous commercial potential because it has modest hardware requirements. “In the end, really, what you need is a halfway-decent projector and a rotating stage,” he says.

### Going big

While chemists work on smarter ways to 3D-print intricate resins, engineers are pushing boundaries in 3D printing of concrete – using computers and robots to precisely automate the pouring process.

The world’s first 3D-printed concrete pedestrian bridge was made by researchers at the Institute for Advanced Architecture of Catalonia in Barcelona, Spain, and installed in a park in Alcobendas, near Madrid, in 2016. Twelve metres long, the bridge features a lattice structure designed with algorithms that maximize strength and reduce the amount of material needed. Other teams have made similar structures, including a 26-metre-long bridge in Shanghai, China, produced by engineers at Tsinghua University in Beijing. And teams and companies in China and the Netherlands have 3D printed demonstration houses.

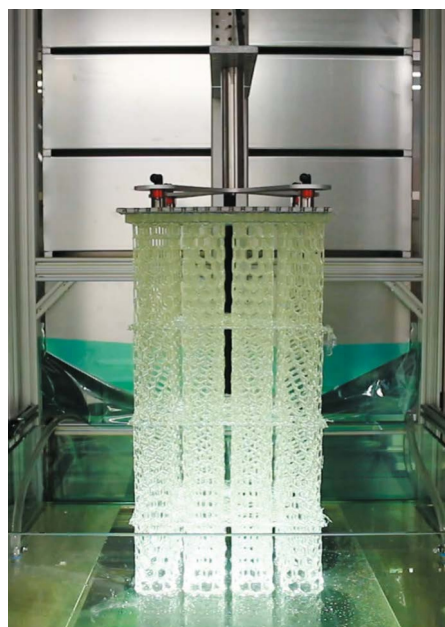
Those structures aren’t constructed in one print job, however: separate segments are printed and then connected. By producing bridges and houses more cheaply and efficiently, 3D printing could reduce concrete’s carbon footprint – but it could also just encourage engineers to build more.

It’s not just concrete that is going big: Amsterdam firm MX3D has printed a bridge from stainless steel. First displayed publicly in 2018, the bridge is now being tested and having sensors installed ahead of a planned installation over an Amsterdam canal. And California start-up firm Relativity Space in Los Angeles says it is constructing a nearly fully

3D-printed rocket. The rocket is designed to lift 1,250 kilograms into low Earth orbit, and its first test launch is slated for the end of this year. Printed metal doesn’t always have the same heat-dissipating performance as non-printed metal, says Relativity Space’s chief executive, Tim Ellis, but the printing process can add cooling channels in geometries that can’t usually be manufactured. Because rockets are used only once or perhaps a few times, they don’t have to be as strong in the long term as do alloys in aeroplane parts, which must resist failure over tens of thousands of pressure cycles, Ellis says.

These large-scale metal-printed projects are built with robot arms that feed a thin metal wire to a laser that welds the material into place. Other established ways to print metal use a laser or a beam of electrons to melt or

**“There are still some big challenges with 4D printing.”**



**A resin printer from Chad Mirkin's lab at Northwestern University in Illinois can create structures as large as a person in hours.**

fuse a bed of powder into layers of finished product. Another technique binds a bed of powder with liquid glue, then sinters the structure in a furnace. And printers designed in the past few years extrude molten metals through nozzles, in much the same way as in FDM.

Aviation firms such as Boeing, Rolls Royce and Pratt & Whitney are using 3D printing to make metal parts, mainly for jet engines. It can be cheaper than milling metal blocks, and the intricate components often weigh less than their conventionally made counterparts.

But 3D-printed metals are prone to defects

that can weaken the final products. Spadaccini and others are trying to use arrays of sensors and high-speed cameras to watch for irregularities such as hotspots of heat or strain – and then make adjustments in real time, he says.

Many scientists are also hoping to improve the intrinsic strength of printed metals, sometimes by controlling the microstructures of the materials. For instance, in October 2017, a US team reported that the intense heat and rapid cooling used in 3D-printing stainless steel could alter the metal’s microstructure such that the product is stronger than those cast conventionally<sup>7</sup>. And two months ago, researchers in Australia and the United States reported a titanium–copper alloy with similar strength advantages<sup>8</sup>. As they solidified, previous 3D-printed titanium alloys tended to form grains that grew in column-like structures. The copper helps to speed up the solidification process, which results in grains that are smaller and sprout in all directions, strengthening the overall structure.

Mark Easton, a materials engineer at RMIT University in Melbourne and one of the leaders of the alloy work, has already had conversations with aerospace companies interested in exploring uses for the material. He says it could also be used in medical implants such as joint replacements.

Many of the printing techniques that work for metals can also be applied to ceramics, with potential applications that include making dental crowns or orthopaedic implants. Moulds for these objects are already made by 3D printing, with the material cast in the conventional way. But 3D-printing the entire object could save time at the dentist or surgeon’s office.

However, it is harder to control the microstructure of 3D-printed ceramics, says Eduardo Saiz, a materials scientist and ceramicist at Imperial College London. And nearly all practical ceramic printing techniques involve extensive post-print sintering that can warp or deform the part. “In my opinion, ceramics is way behind polymers and metals in terms of practical applications,” he says.

### Change over time

The field’s future could also lie in ‘4D printing’ – 3D-printed objects that also have the ability to perform some mechanical action, akin to artificial muscles. Often, these incorporate shape-memory polymers, materials that can react to changes in their environment such as heat or moisture.

In May 2018, researchers at the Swiss Federal Institute of Technology (ETH) in Zurich and the California Institute of Technology in Pasadena reported printing a submarine that propels itself forward using paddles that snap backwards when placed in warm water<sup>9</sup>. The work could lead to microrobots that can explore the oceans autonomously. But for the

NORTHWESTERN UNIV.



These multi-material print heads can switch between printing hard and soft materials in one object.

moment, the paddles must be reset after each stroke. Such devices could use battery power to reset themselves, but that makes the machine less efficient than one made conventionally, says Geoff Spinks, a materials engineer at the University of Wollongong in Australia. “There are still some big challenges with 4D printing,” he says.

Another approach to 4D-printed devices involves triggering the action with a changing external magnetic field. US researchers have 3D-printed lattice structures filled with a liquid that changes stiffness in response to a magnetic field<sup>10</sup> – which could perhaps be used to help car seats stiffen on impact.

Other, more passive potential 4D printing applications include stents, which could be compressed to be implanted then expanded on reaching the desired site in a blood vessel to prop it open. Last July, researchers in Switzerland and Italy described a 4D-printed stent that is just 50 micrometres wide<sup>11</sup>, much smaller than conventional ones. The devices are so small, the team says, they could one day be used to treat complications in fetuses, such as strictures in the urinary tract, which can sometimes be fatal.

Perhaps the most ambitious example of 4D printing is matter that not only moves, but is alive. Currently, techniques for such bioprinting can print tissue, such as human skin, that is suitable for lab research, as well as patches of tissue for livers and other organs that have been successfully implanted in rats. But such techniques are still far from ready to integrate into a human body. Researchers dream of printing fully functioning organs

that could alleviate long wait lists for organ donors. “I personally feel we’re a decade-plus away from that, at least, if ever,” says Lewis.

### All together now

Many inventive ideas about printing matter that moves or changes rely on printing multiple materials together. “That’s absolutely where the field is heading,” says Scott.

Last November, Lewis and her lab described a printer that can rapidly switch between different polymer inks or mix them as it prints a single object<sup>12</sup>. This means objects can be printed with both flexible and rigid parts. Lewis has spun off previous work on multi-material printers into a firm called Voxel8, a start-up in Somerville, Massachusetts. Her multi-material printer could help with the athletics wear that Voxel8 is developing, says Lewis. Wearable devices need to be flexible around joints while also having rigid parts to house electronics. Saiz calls the printer “beautiful work”, adding wistfully: “There’s nothing like that for ceramics or metal.”

And in March 2018, a team led by Jerry Qi, a materials engineer at Georgia Institute of Technology in Atlanta, unveiled a four-in-one printer. This combines a nozzle that extrudes molten polymer with one that prints light-sensitive resin, ready to be cured by ultraviolet lamps or lasers, and two that print wires and circuitry from tiny dots of metal<sup>13</sup>. The print heads work together to make integrated devices with circuits embedded on a rigid board or inside a flexible polymer enclosure. Qi says his group is now collaborating with electronics companies interested in printing circuit-board prototypes

faster than conventional methods.

It wasn’t as simple as bolting four different printers into one platform: the researchers also needed to develop software that would allow each print head to communicate with the others and keep track of the progress.

The field is still far from delivering on early visions of bringing mass manufacturing into people’s homes. For now, sophisticated printers are too expensive to appeal to non-specialists. But 3D printing has come a long way in the past 20 years. Todd remembers people touring his lab in the early 2000s to see his technique to fuse specks of metal dust together to grow parts. Compared with the conventional milling machines and metal-cutting systems in neighbouring labs, his 3D-printing machines struck visitors as a complete oddity. “It was like we were some sort of a dog playing a piano in a bar,” he recalls. Now, for many firms, that trick is standard practice.

**Mark Zastrow** is a writer based in Seoul.

1. Walker, D. A., Hedrick, J. L. & Mirkin, C. A. *Science* **366**, 360–364 (2019).
2. Hull, C. W. Apparatus for production of three-dimensional objects by stereolithography. US patent 4575330A (1984).
3. Tumbleston, J. R. et al. *Science* **347**, 1349–1352 (2015).
4. de Beer, M. P. et al. *Sci. Adv.* **5**, eaau8723 (2019).
5. Kelly, B. E. et al. *Science* **363**, 1075–1079 (2019).
6. Loterie, D., Deirrot, P. & Moser, C. Preprint at ResearchGate <https://doi.org/10.13140/RG.2.2.20027.46889> (2018).
7. Wang, Y. M. et al. *Nature Mater.* **17**, 63–71 (2018).
8. Zhang, D. et al. *Nature* **576**, 91–95 (2019).
9. Chen, T., Bilal, O. R., Shea, K. & Daraio, C. *Proc. Natl Acad. Sci. USA* **115**, 5698–5702 (2018).
10. Jackson, J. A. et al. *Sci. Adv.* **4**, eaau6419 (2018).
11. de Marco, C. et al. *Adv. Mater. Technol.* **4**, 1900332 (2019).
12. Skylar-Scott, M. A., Mueller, J., Visser, C. W. & Lewis, J. A. *Nature* **575**, 330–335 (2019).
13. Roach, J. D. et al. *Add. Manuf.* **29**, 100819 (2019).





ILLUSTRATIONS BY JOANNA GEBAL

# A NEW TWIST ON GENE EDITING

As an alternative to CRISPR, RNA editing could offer flexible, reversible therapies.  
By Sara Reardon

**T**horsten Stafforst found his big break at the worst possible time. In 2012, his team at the University of Tübingen in Germany discovered that by linking enzymes to engineered strands of RNA, they could change the sequences of messenger RNA molecules in cells. In essence, they could rewrite the genome's instructions en route to making proteins.

The process could theoretically serve to treat numerous diseases, both ones with genetic underpinnings and those that would benefit from a change in the amount or type of a protein being produced. But Stafforst had a lot of trouble getting the discovery published – it was simply not interesting any more. His finding<sup>1</sup> was

overshadowed by the discovery a few months earlier that the DNA-editing tool CRISPR–Cas9 could be used to permanently alter the genome.

Since then, CRISPR has become a fixture in the laboratory and has spawned a number of companies aimed at using the technology to develop drugs and treatments. With CRISPR sucking up all the attention, Stafforst says, people reacted to his paper with indifference. They asked, “Why do we need this when there’s DNA editing?”

But CRISPR editing – at least as a therapeutic technique in people – has turned out to be more difficult than initially thought. Researchers have documented ways that Cas9, one of the enzymes used in CRISPR gene editing, could trigger immune responses, or cause

accidental changes to the genome that would be permanent. RNA editing, by contrast, could allow clinicians to make temporary fixes that eliminate mutations in proteins, halt their production or change the way that they work in specific organs and tissues. Because cells quickly degrade unused RNAs, any errors introduced by a therapy would be washed out, rather than staying with a person forever.

Excitement over RNA editing is finally catching on. In 2019, researchers published more than 400 papers on the topic, according to data from Scopus, an abstract and citation database. A handful of start-up companies are beginning to use RNA-editing systems to develop potential treatments for everything from genetic diseases such as muscular dystrophy to temporary maladies such as acute pain. And although RNA-based drugs have had difficulty reaching the market owing to challenges in delivery and tolerance, some regulatory approvals in the past few years might help to pave the way for RNA-editing therapies.

Several hurdles remain: current technologies can alter RNA sequences in only a few limited ways, and getting the system to work as intended in the human body will prove challenging. Still, researchers hope that new technologies, such as protein engineering, and improved methods for delivering RNA to cells can help to overcome these limitations. “It really opens a world we haven’t seen before,” Stafforst says.

## A role for RNA

A foundational tenet in molecular genetics – its central dogma – was that cellular machinery faithfully transcribes genetic information from a double-stranded DNA template into a single-stranded RNA messenger, which is then translated into a protein. But in the 1980s, a handful of labs noticed that some mRNA transcripts contained altered or extra letters that were not encoded in the DNA. The findings were controversial until scientists uncovered a family of enzymes called adenosine deaminases acting on RNA (ADARs). These proteins bind to RNAs and alter their sequence by changing a familiar base known as adenosine into a molecule called inosine. Although not one of the canonical RNA bases, inosine is read by the cell’s protein-translation machinery as the familiar guanosine. A handful of other RNA-editing enzymes surfaced around the same time.

Scientists have struggled over the past three decades to understand what exactly RNA editing accomplishes. The editors work only on double-stranded RNAs, which sometimes show up in the cell as regulatory elements – or as viruses. Some have speculated that the ADAR proteins evolved as a defence against viruses, but many viruses with double-stranded RNA are unaffected by the enzymes. The editing might serve a regulatory function, but most adult tissues don’t produce the high levels of



the proteins required for the editing to occur.

Brenda Bass, a biochemist at the University of Utah in Salt Lake City, was among the first to identify ADARs in frog embryos<sup>2</sup>. She says that no one has found a specific role for the changes made to non-protein-coding RNAs, which account for the majority of edited molecules. The editing could serve to protect double-stranded RNAs from immune attack. Bass suspects that ADARs edit the double-stranded transcripts, adding inosines as a way of telling the body to leave them alone. The enzymes also seem to have a role in embryonic development: mice that lack ADAR genes die before birth or don't live long after. The editors also seem to have some function in select tissues of adult organisms – such as the nervous system of cephalopods.

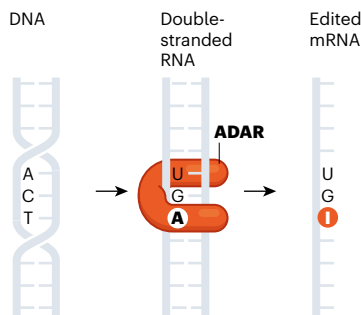
It was this activity that drew marine biologist Joshua Rosenthal to RNA editing in the early 2000s. It seems that highly intelligent cephalopods, such as squid, cuttlefish and octopuses, use RNA editing extensively to adjust genes involved in nerve-cell development and signal transmission. No other animals are known to use RNA editing in this way. Inspired by these observations, Rosenthal wondered whether it was possible to use the system to correct the messages produced by dysfunctional genes in a therapeutic setting. In 2013, his group at the University of Puerto Rico in San Juan re-engineered ADAR enzymes and attached them to guide RNAs that would bind to a specific point in an mRNA – creating a double strand. With these, they were able to edit transcripts in frog embryos, and even in human cells in culture<sup>3</sup>.

Similar to Stafforst, Rosenthal, now at the Marine Biological Laboratory in Woods Hole, Massachusetts, saw his publication mostly ignored. A similar fate, he learnt, had befallen the work of researchers at a company called Ribozyme, who in 1995 proposed 'therapeutic editing' of mutated RNA sequences by inserting complementary sequences into frog embryos and allowing ADARs to edit the resulting double-stranded molecule and correct the mutation<sup>4</sup>.

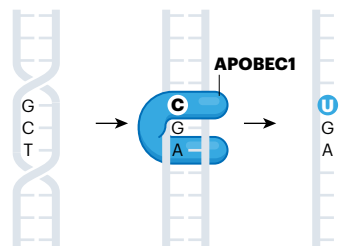
But in the past several years, multiple factors have converged to bring Rosenthal's and Stafforst's findings to the fore. Peter Beal, a chemist at the University of California, Davis, says that the 2016 publication<sup>5</sup> of the molecular structure of ADAR bound to double-stranded RNA made the system more understandable and enabled scientists to better engineer the enzyme to enhance its delivery or make it more efficient. And in 2018, the US Food and Drug Administration (FDA) approved the first therapy using RNA interference (RNAi): a technique in which a small piece of RNA is inserted into a cell in which it binds to native mRNAs and hastens their degradation. The approval has opened the door for other therapies that involve mRNA interactions, says Gerard Platenburg, chief innovation officer

## THE RNA CORRECTIONS

Several RNA-editing proteins can work on double-stranded RNAs, including the messenger RNAs that encode proteins. They can make very specific single-letter changes: an enzyme called ADAR changes the base known as adenosine into an inosine molecule (which protein-producing machinery reads as a guanosine).



A protein called APOBEC1 can change cytosine into uracil. Researchers can use this enzyme to change the sequence of the resulting protein or add and remove sequences that prematurely halt the production of a protein.



of ProQR Therapeutics in Leiden, the Netherlands, which is pursuing various RNA-based therapies. "Learning from the past, and with the number of approvals picking up, the field has matured a lot," says Platenburg.

Many see RNA editing as an important alternative to DNA editing using techniques such as CRISPR. CRISPR technology is improving, but DNA editing can cause unwanted mutations in other parts of the genome – 'off-target effects' – which might create new problems.

Rosenthal expects, moreover, that RNA editing will prove useful for diseases without a genetic origin. He is currently using ADARs to edit the mRNA for a gene encoding the sodium channel Nav1.7, which controls how pain signals are transmitted to the brain. Permanently changing the Nav1.7 gene through DNA editing could eliminate the ability to feel pain and disrupt other necessary functions of the protein in the nervous system, but tuning it down through RNA editing in select tissues for a limited amount of time could help to alleviate pain without the risk of dependency or addiction associated with conventional painkillers.

Similarly, RNA editing could allow researchers to mimic genetic variants that provide a health advantage. For example, people with certain mutations in the gene *PCSK9*, which regulates cholesterol in the bloodstream, tend to have lower cholesterol levels, and modifying *PCSK9* mRNA could confer a similar advantage

without permanently disrupting the protein's other functions. Immunologist Nina Papavasiliou of the German Cancer Research Center in Heidelberg says that RNA editing could be used to fight tumours. Some cancers hijack important cell-signalling pathways, such as those involved in cell death or proliferation. If RNA editors could be conscripted to turn off key signalling molecules temporarily, she says, "we could see the tumour die". Then, the patient could stop the therapy, allowing the pathway to resume its normal functions.

As a treatment, RNA editing might be less likely to cause a potentially dangerous immune reaction than are CRISPR-based approaches. Unlike the DNA-editing enzyme Cas9, which comes from bacteria, ADARs are human proteins that don't trigger an attack from the immune system. "You really don't need heavy machinery to target RNA," says Prashant Mali, a bioengineer at the University of California, San Diego.

In a paper published last year<sup>6</sup>, Mali and his colleagues injected guide RNAs into mice born with a genetic mutation that causes muscular dystrophy. The guide RNAs were designed to trigger production of a missing protein called dystrophin. Although the system edited only a small amount of the RNA encoding dystrophin, it restored the protein to about 5% of its normal level in the animals' muscle tissue, an amount that has shown therapeutic potential.

In other diseases that result from a missing or dysfunctional protein, such as some types of haemophilia, "it makes a huge difference to go from nothing to something", Stafforst says, and it might not be necessary to edit RNA in every cell in the body. RNA editing might perform better than forms of gene therapy that would involve injecting a new gene. Mali and others say that directing native ADARs to operate on the cell's own mRNA might provide a more natural response than introducing an external, engineered gene.

RNA-editing technology is far from perfect, however, even when it comes to laboratory applications. "It is early days," Bass says. "There's lots of questions." Because ADARs are much less efficient than CRISPR, they could be less useful for making genetically modified plants and animals. "As a research tool, it's very limiting," says Jin Billy Li, a geneticist at Stanford University in California.

Another major disadvantage is that ADARs can make only a few kinds of change to RNA. CRISPR systems act as scissors by cutting DNA at a designated spot and removing or inserting a new sequence; ADARs are more like an overwrite function that changes letters chemically, without breaking the RNA molecule's 'backbone'.

Although this process is less likely to cause unintended mutations, it limits the enzymes to making specific changes – adenosine to inosine in the case of ADARs, and cytosine to



uridine by a set of enzymes called APOBECs (see ‘The RNA corrections’). There are a few other possibilities. Grape plants, for instance, can change cytidines to uridines, and some tumours can change guanines to adenosines. “Biodiversity is giving us tons of answers to these things,” Rosenthal says. “I think down the line, things like the squid are going to teach us a lot.” But he says the field is understudied — researchers don’t understand the process that drives this editing. And it remains to be seen whether a plant enzyme, for instance, could function in human cells.

Scientists are already looking for ways to engineer new enzymes that could expand RNA-editing capabilities. “It’s quite a process where you don’t know what you’ll find,” says Omar Abudayyeh, a biological engineer at the Massachusetts Institute of Technology (MIT) in Cambridge. Working with Feng Zhang, a CRISPR pioneer at MIT, Abudayyeh and his colleagues linked an ADAR enzyme to Cas13 (ref. 7). A bacterial enzyme similar to the CRISPR-associated protein Cas9, Cas13 cuts RNA instead of DNA. The researchers altered the sequence of the ADAR until it could convert cytidines to uridines. They then used the new system in human cells to change bases in mRNAs encoded by several genes, including *APOE*. One naturally occurring genetic variant of this gene is associated with Alzheimer’s disease, and editing it could switch the variant to the harmless form.

Abudayyeh and his MIT collaborator, biological engineer Jonathan Gootenberg, admit it is possible that changing the ADAR protein could cause the immune system to stop recognizing it as a natural human protein and

attack cells that contain it. But they say that because these edits are small, this risk pales next to known concerns about the immune system attacking Cas13 or the virus used to deliver the editing tools into cells.

Researchers see promise in a natural process called pseudouridylation, in which a set of protein and RNA enzymes chemically modify the structure of uridines in mRNA. Unlike ADAR modifications, pseudouridylation doesn’t change the sequence of the mRNA or protein. Instead, for reasons that are not entirely clear, the process stabilizes the RNA molecule and causes the translation machinery to ignore signals instructing it to stop making protein.

The ability to turn these molecular red lights into green lights could be powerful. Yi-Tao Yu, a biochemist at the University of Rochester in New York, says that hundreds of genetic diseases are caused by DNA mutations that create incorrect stop signals in mRNAs, resulting in a shortened protein that doesn’t function normally in the body. “The list is very long,” Yu says, and includes cystic fibrosis, the eye disease Hurler’s syndrome and numerous cancers.

Despite its early stage, researchers — and biotech investors — are excited about the wide potential of RNA editing. “I got into it way before it became cool,” says Papavasiliou, who is trying to map where natural ADARs work in the body. “For many years this was a backwater, and all of a sudden there’s a company popping up every two weeks.”

Numerous start-ups and established DNA-editing firms have announced their intention to move into RNA. They include Beam Therapeutics in Boston, Massachusetts, which was co-founded by Zhang and Liu and has been

developing CRISPR DNA editing as a therapy for several blood diseases. Locana, based in San Diego, is also pursuing CRISPR-based RNA editing that it hopes could treat conditions including motor-neuron disease and Huntington’s disease.

The challenge for industry is to work out the best way to get the guide RNAs into the cell without triggering an immune reaction or causing the cell to degrade them. Beal says that this could include making strategic chemical modifications to the engineered RNAs that stabilize them, or embedding them in a nanoparticle or virus that can sneak into cells.

And although ADARs are already in human cells, the human body makes only small amounts of them in most tissues, meaning that any therapy might need to add ADARs or other enzymes to boost cells’ editing capabilities. Packing viruses with the genes that encode all the machinery needed for RNA editing might not be efficient. Many hope that it won’t be necessary.

Platenburg hopes to add RNAs and rely on the naturally occurring ADARs to help to correct the lettering of mRNAs that contribute to retinal disorders. “We use the system given to us by nature and harness it,” he says.

Researchers including Stafforst are engineering guide RNAs with chemical modifications that attract ADARs in the cell to the editing site. But some researchers worry that conscripting the natural ADARs into editing specific mRNAs could pull them away from their normal tasks and cause other health problems. Altering gene expression in one part of the body could affect other parts in unforeseen ways. In Mali’s muscular-dystrophy study, for instance, mice developed liver problems for unknown reasons. “It’s a tool in development still,” he says.

“ADAR evolved to allow the body to modify bases in a very targeted fashion,” says Nessim Bermingham, chief executive and a co-founder with Rosenthal and others of biotechnology company Korro Bio in Cambridge, Massachusetts. Bermingham is optimistic about the prospects of RNA editing, but cautious not to get ahead of the biology. “We have a lot of work to do as we start to mature these techniques,” he says. “We’re not leaving anything off the table, but we have to recognize certain limitations.”

**Sara Reardon** is a freelance journalist in Bozeman, Montana.

- Stafforst, T. & Schneider, M. F. *Angew. Chem. Int. Ed. Engl.* **51**, 11166–11169 (2012).
- Bass, B. L. & Weintraub, H. *Cell* **55**, 1089–1098 (1988).
- Montiel-Gonzalez, M. F., Vallecillo-Viejo, I., Yudowski, G. A. & Rosenthal, J. J. C. *Proc. Natl. Acad. Sci. USA* **110**, 18285–18290 (2013).
- Wolff, T. M., Chase, J. M. & Stinchcomb, D. T. *Proc. Natl. Acad. Sci. USA* **92**, 8298–8302 (1995).
- Matthews, M. M. et al. *Nature Struct. Mol. Biol.* **23**, 426–433 (2016).
- Katrekar, D. et al. *Nature Methods* **16**, 239–242 (2019).
- Abudayyeh, O. O. et al. *Science* **365**, 382–386 (2019).



# Books & arts



JAMES MACDONALD/BLOOMBERG VIA GETTY

**Sugar:** multinationals have used the tobacco-industry playbook to stymie legislation aimed at cutting consumption.

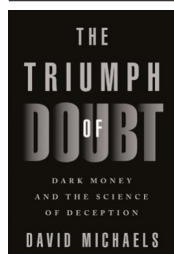
## Truth decay: when uncertainty is weaponized

From tobacco to food and fuels, industries use denial, deceit and doubt to corrupt. **By Felicity Lawrence**

In 2017, US presidential strategist Kellyanne Conway coined the phrase “alternative facts” to defend false claims about the size of the crowd at Donald Trump’s inauguration. Numerous commentators lamented that we were entering a new era of Orwellian doublethink.

These are indeed upside-down times, as epidemiologist and former safety regulator David Michaels demonstrates in his excoriating account of the corporate denial industry, *The Triumph of Doubt*. Unwelcome news is automatically rebranded fake news. Inconvenient evidence from independent sources – say,

about climate breakdown and fossil fuels, or air pollution and diesel emissions – is labelled junk science and countered with rigged studies claiming to be sound.



**The Triumph of Doubt: Dark Money and the Science of Deception**  
David Michaels  
Oxford Univ. Press  
(2020)

But it would be wrong to see truth decay solely as the preserve of today’s populist politicians. Normalizing the production of alternative facts is a project long in the making. Consultancy firms that specialize in defending products from tobacco to industrial chemicals that harm the public and the environment have made a profession of undermining truth for decades. They hire mercenary scientists to fulfil a crucial role as accessories to their misrepresentations.

### Denial machine

Michaels was among the first scientists to identify this denial machine, in his 2008 book *Doubt is Their Product*. His latest work combines an authoritative synthesis of research on the denial machine published since then with his own new insights gleaned from battles to control the toxic effects of a range of substances. He takes on per- and poly-fluoroalkyls, widely used in non-stick coatings, textiles and firefighting foams; the harmful effects of alcohol and sugar; the disputed role of the ubiquitous glyphosate-based pesticides in cancer; and the deadly epidemic of addiction to prescribed opioid painkillers. In each

case, Michaels records how the relevant industry has used a toolbox of methods to downplay the risks of its products, spreading disinformation here, hiding evidence of harm there, undermining authorities – all tactics from the tobacco industry's playbook.

The doubt in the title of both Michaels's books derives from a now-notorious memo written in 1969 by an unnamed executive at a subsidiary of British American Tobacco. It outlined a strategy for maintaining cigarette sales: "Doubt is our product since it is the best means of competing with the 'body of fact' that exists in the minds of the general public. It is also the means of establishing a controversy." By creating scientific disinformation about links between tobacco and disease, this malign strategy delayed regulation by decades and protected corporate profits.

Michaels's insider perspective on the doubt machine dates back to 1998, when he became chief safety officer for nuclear-weapons facilities at the US Department of Energy during the administration of US president Bill Clinton. Here, he had a ringside view of the tricks used by vested interests to dispute established science, intimidate the authorities and scupper regulation. In his first book, he described how the 'product defence industry' applied the tobacco template to asbestos, lead, plastics and toxic materials such as beryllium used in nuclear applications.

From 2009 to 2017, Michaels served as a senior regulator, appointed by president Barack Obama, in the Occupational Safety and Health Administration (OSHA). Here, he gathered even more material to show how deceptions have infected the body politic.

### Subverting the method

The principles of scientific inquiry involve testing a hypothesis by exploring uncertainty around it until there is a sufficient weight of evidence to reach a reasonable conclusion. Proof can be much longer in coming, and consensus still longer. The product-defence industry subverts these principles, weaponizing the uncertainty inherent in the process. Its tricks include stressing dissent where little remains, cherry-picking data, reanalysing results to reach different conclusions and hiring people prepared to rig methodologies to produce funders' desired results.

Michaels acknowledges other doubt scholarship. This includes that of science historians Naomi Oreskes and Erik Conway in *Merchants of Doubt* (2010); nutritional scientist Marion Nestle's numerous books on the food industry, such as *Soda Politics* (2015)



A man prays in foam caused by pollution in the Yamuna River in New Delhi.

and *Unsavory Truth* (2018); and journalist Jane Mayer's 2016 *Dark Money*. That last book traced the funding that links climate-change denial to the libertarian right's ideological drive to shrink the state and deregulate industry.

Michaels names names fearlessly, pointing the finger at product-defence practitioners and the front groups and think tanks that masquerade as independent while taking

**“Creating scientific disinformation delayed regulation by decades and protected corporate profits.”**

industry's shilling. Those wanting to check his allegations can find many previously unavailable source documents archived at the Triumph of Doubt Special Collection at <https://toxicdocs.org>.

### Emissions cheats

So much of his material outrages, but two episodes stand out. One is the German car manufacturer Volkswagen's brazen malfeasance regarding its diesel engines. The company developed secret software so these engines could cheat emissions tests, allowing its vehicles to fraudulently pass stringent US checks on the disease-causing particulates in diesel exhaust. This was unintentionally uncovered in 2014 by students working on behalf of the campaign group the International Council on Clean Transportation in Washington DC,

and confirmed the following year by the US Environmental Protection Agency. Michaels's account of the scientists prepared to launder their data to abet this criminal activity is forensic.

The second standout is his description of his years-long battle at the OSHA to reduce workers' exposure to silica particles from sand used in dozens of industries, from construction to steel manufacture and fracking. He and at least 50 staff members worked to collate evidence and counter a barrage of pseudoscientific objections and litigation.

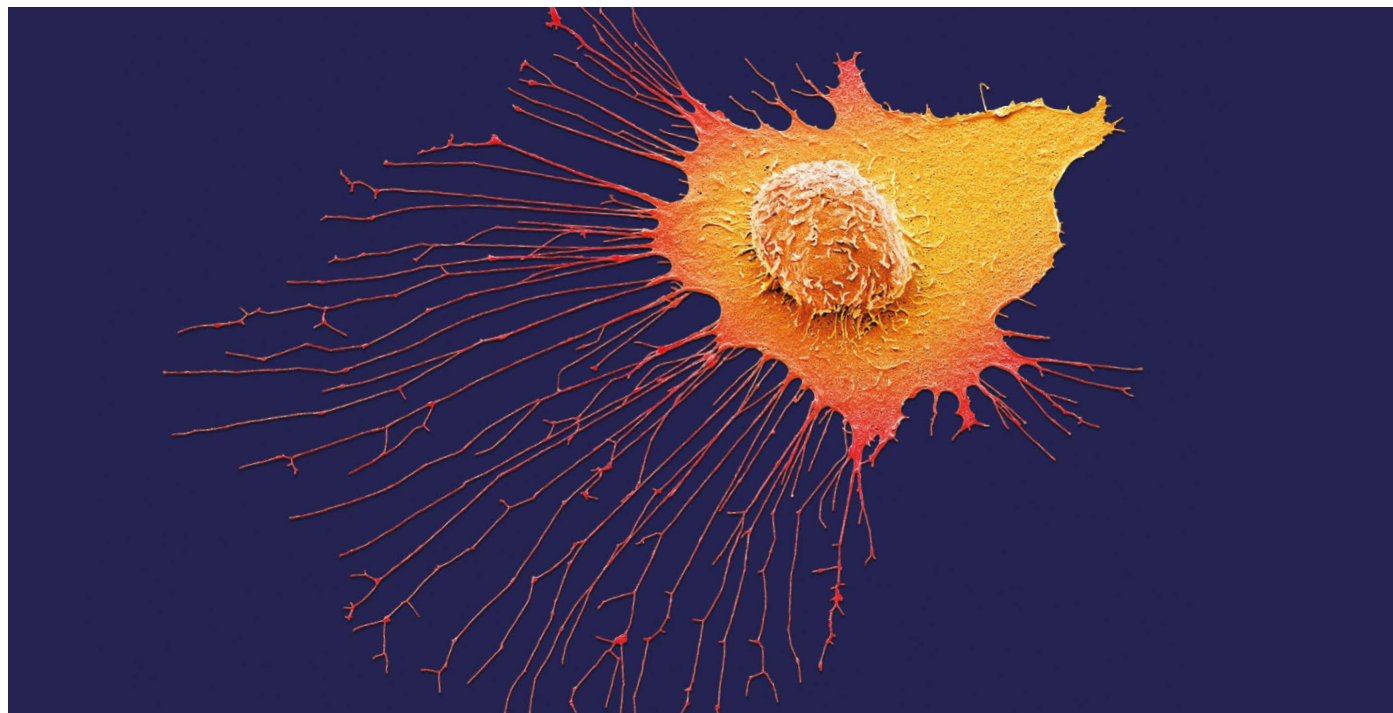
Michaels, no longer required to be a non-partisan government official, reserves special criticism for the Republican Party. He argues that corporate polluters and manufacturers of dangerous products have long depended on the party to neuter public-health and regulatory agencies with phoney rhetoric about liberty and free-market enterprise. He wants stronger regulation, not because he does not care about freedom, he says, but because we cannot be free without the state's protection from harm.

*The Triumph of Doubt* is at times dense with technical detail, of necessity as Michaels prosecutes his case against companies known to be litigious. It is a brave and important book, raising the alarm about the systemic corruption of science.

**Felicity Lawrence** is special correspondent for *The Guardian* in London and author of *Not on the Label* and *Eat Your Heart Out*.  
e-mail: [felicity.lawrence@theguardian.com](mailto:felicity.lawrence@theguardian.com)



# Comment



A migrating breast cancer cell.

## Genomics: data sharing needs an international code of conduct

Mark Phillips, Fruzsina Molnár-Gábor, Jan O. Korbel, Adrian Thorogood, Yann Joly, Don Chalmers, David Townsend & Bartha M. Knoppers for the PCAWG Consortium.

**Efforts to protect people's privacy in a massive international cancer project offer lessons for data sharing.**

**M**ore than 800 terabytes of genomic data are available to investigators all over the world, thanks to a major international project to identify the genetic traits associated with various types of cancer. Researchers involved have just published six papers in *Nature*. (Another 16 papers have been published elsewhere.)

All eight of us were involved in the six-year endeavour. And four of us helped put in place safeguards to protect the privacy of the thousands of patients and volunteers

who consented to have their data used in the research. Here, we reflect on some of the lessons learnt for researchers sharing vast amounts of genomic data.

Genomics researchers worldwide are increasingly dealing with vast data sets gathered by consortia spanning many countries. Most are unclear on what to do to protect people's privacy and to comply with international and national data-protection laws, especially given recent and ongoing changes in legislation.

An international code of conduct for genomic data is now crucial. Built by the genomics community, it could be updated as technologies and knowledge evolve more easily than is possible for national and international legislation.

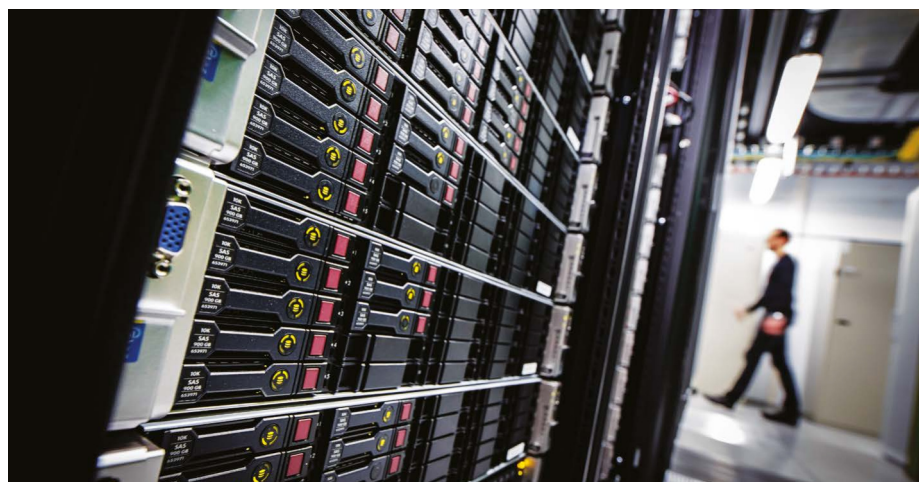
### In the clouds

Between 2013 and 2019, 468 institutions from 34 countries in Asia, Australasia, Europe

and North America amassed 2,658 cancer genomes – each paired with a non-cancerous sequence from the same person. The effort was led by the International Cancer Genome Consortium (ICGC).

The combined data were made available to investigators largely thanks to cloud computing. The project – the Pan-Cancer Analysis of Whole Genomes (PCAWG) – is the first to try to aggregate so many subprojects across different jurisdictions and make the entire data set available across the world.

Much of the PCAWG data (and the tools for analysing them) were made available through the Cancer Genome Collaboratory, a cloud service built for the genomic research community. (The commercial cloud-service provider Amazon Web Services was also used.) But the data were first processed in high-performance computer centres and the clouds of academic



Server racks in a centre in Berlin.

institutions in Germany, the United Kingdom, the United States, Canada, Spain, Japan and South Korea. Some were also processed using commercial clouds (Amazon Web Services, Microsoft Azure and Seven Bridges).

Since PCAWG began, several other international genomics projects have turned to the cloud (see ‘Cashing in on clouds’) including The Human Cell Atlas, an international project to create a reference map of all human cells<sup>1</sup>, and the European Open Science Cloud, which is for researchers and professionals in science, technology, the humanities and social sciences<sup>2</sup>.

This trend is likely to continue. Cloud services are becoming cheaper and more readily available, and researchers are increasingly reaping the benefits of sharing ever-larger amounts of genomic data with international colleagues<sup>3</sup> (see ‘Open data’).

Yet cloud services bring fresh challenges with respect to the protection of participants’ data – especially given that national governments, law enforcement and private corporations are increasingly showing interest in accessing them. Canadian border authorities, for example, are choosing which country to deport migrants to on the basis of DNA test results from consumer genomic services<sup>4</sup>.

## Long-term continuity

Organizations such as the Cancer Genome Collaboratory can persist only for as long as they are funded. Even in the case of Amazon and other major tech companies, service outages caused by technical problems, changes to the company’s terms of service or even sudden closure of the company could block researchers’ access to data at any time. Also, it is often unclear to what extent researchers using cloud services can ensure that their data are not disclosed to third parties, such as those conducting abusive state-level surveillance. Nor is it clear what steps must be taken to protect the data against such breaches of confidentiality.

In the case of PCAWG, the ICGC’s Data Access

Compliance Office helped to guard against some of these issues. Anyone wanting to use PCAWG data entered into a contract with the project’s data-access committee; they had to confirm, for instance, that they would not try to re-identify patients or volunteers once these people’s data had been stripped of personal information. No breach of donor confidentiality is known to have occurred.

But even when researchers request the data from an associated data-access committee (as in the case of PCAWG and elsewhere), numerous issues remain unresolved. It is unclear, for instance, what vetting should

## Cashing in on clouds

**Cloud services have been transformative in enabling large-scale genomic analysis.**

Conventionally, any research team wanting to analyse an aggregate data set collected by a consortium would first have to seek authorization from each project partner’s research ethics or data-access compliance office. It would then have to download the data from each subproject over the Internet or — more likely — have the hard drives containing the data sent by post. In the case of the Pan-Cancer Analysis of Whole Genomes, which comprises 800 terabytes of raw data, investigators were able to save months, and thousands of dollars, by immediately accessing the data they needed, and experimenting with and customizing the analytical tools developed by the community. They could also obtain authorization to access most of the data from one place — the International Cancer Genome Consortium’s Data Access Compliance Office. **M.P. et al.**

occur before researchers get access to sensitive genomic data, or what checks should be made before genomic data are shared internationally. Even those involved in PCAWG could not establish a truly international cloud because of restrictions on the transfer of data across borders (caused, in this case, by European regulators having concerns about genomic data from Europeans being held in the United States).

The US component of the project (The Cancer Genome Atlas; TCGA), which contributed one-third of all PCAWG samples, was made available to researchers through the University of Chicago’s Protected Data Cloud. Researchers wanting to obtain those data had to abide by an access agreement that was largely compatible with that provided by the ICGC’s Data Access Compliance Office. Ultimately, however, TCGA remained conceptually split off from the rest of the project, because researchers had to follow a different access procedure and to combine the two data sets themselves.

## Code of conduct

Genomics researchers urgently need clear data-sharing rules that are harmonized across jurisdictions.

An international code of conduct could help investigators to overcome some of the current hurdles, as well as others that might arise as legislation on data protection evolves.

Such a code could outline the steps researchers must take to comply with the European Union’s General Data Protection Regulation (GDPR), implemented in 2018, and the US Health Insurance Portability and Accountability Act, among other laws. In fact, the GDPR explicitly encourages the development of sector-specific data-protection codes in its Article 40. And last June, the European Data Protection Board (an independent body tasked with issuing guidance on the GDPR and encouraging the drawing-up of codes of conduct), issued guidelines on the submission, approval and monitoring of such codes for data processing. It also promised further guidance on the use of codes as a potential way to facilitate the transfer of data across borders (see [go.nature.com/322nkv](https://go.nature.com/322nkv)).

A European biobanking research infrastructure, known as BBMRI-ERIC, announced in 2017 that it would develop an EU-wide Code of Conduct on Health-Related Data, to submit to the European Commission (see [go.nature.com/2j3ihce](https://go.nature.com/2j3ihce)). When completed, and if approved, such a European code could be beneficial. Meanwhile, we call on the genomics research community to prioritize the establishment of an international code of conduct that lays out how existing ethical and legal obligations can be satisfied in relation to international genomic clouds.

At least five aspects must be considered.



**Identifiability.** Despite the problems with it<sup>5</sup>, de-identification, in which health data are stripped of any information that could be used to identify the participant (such as name, social security number, address) has long been hailed as a way to protect people's privacy in research<sup>6</sup>. Yet because of conflicting terminology and gaps in understanding, researchers rarely know what standard they must meet for their data to be properly anonymized or 'pseudonymized' (in which a code enables individuals to be re-identified)<sup>7</sup>. What's more, laws are difficult to enforce in practice because it is often unclear how breaches of confidentiality occurred, or which organization or researcher was responsible<sup>8,9</sup>.

Data-protection laws, such as the GDPR or the California Consumer Privacy Act, invariably require the identifiability of data to be analysed on a case-by-case basis, in part because the technological tools enabling identification are constantly changing. Even though it is difficult to lay out hard and fast rules in advance, a code could provide some guidance on how to evaluate when it is reasonable to deposit genomic (and health data more broadly) in open-access repositories. This might involve considering, say, whether a set of genomic variants is somatic or present in the germline and so inherited. (Researchers have shown that it is possible to identify an individual using only a few germline variants<sup>10</sup>; no one has yet been able to identify someone on the basis of somatic tumour variants.)

**Broad consent.** The GDPR explicitly recognizes an exception to 'specific consent', meaning the consent people give for their data to be used in a specific research project. This is to allow participants' data to be used for certain areas of scientific research, in keeping with recognized ethical standards<sup>11</sup>. Guidance is needed on what researchers must do to meet the requirements for broad consent. Furthermore, how should they keep patients and volunteers informed about how their data are ultimately used?

**Return of individual findings, portability and access.** How to safeguard participants' right to move their data around – by giving them their data in a machine-readable format, rather than as a printed PDF, for example – should be clarified. The code could also lay out what steps are necessary for responsible communication of health data to a patient or volunteer<sup>12,13</sup>. Should people who are being informed about the identification of genomic variants of malignant or unknown significance be offered genetic counselling, for instance?

**Withdrawal.** Researchers need guidance on how they can meet participants' right to withdraw from research. The GDPR requires that those entrusted with people's data keep

## Open data

**Over the past three decades, geneticists around the world have been sharing more and more data.**

Last year, more than 83,000 researchers from 146 countries downloaded 6.7 petabytes of (mainly human) DNA data from the European Molecular Biology Laboratory's European Bioinformatics Institute. This hosts many biological data sets and makes them accessible worldwide. That is equivalent to around 230 billion whole human genomes.

Such sharing of genomic data will only increase as more data become available. By 2025, more than 60 million patients worldwide are expected to have had their genome or exome (protein-coding regions) sequenced as part of routine health care<sup>16</sup> – potentially providing a formidable resource for researchers. **M.P. et al.**

records of third parties to whom they have disclosed those data. And when consent is revoked, they must notify the third parties. Yet all sorts of questions remain, such as whether analyses on aggregate data should be revised with the participants' data removed, and so on.

**Compelled disclosure.** A code of conduct could provide researchers with guidance on how to deal with government requests for personal data, including what legal protections

## "Researchers need guidance on how they can meet participants' right to withdraw from research."

they can appeal to. In the United States, for example, the National Institutes of Health's Certificates of Confidentiality are designed to shield researchers from such requests<sup>14</sup>.

### Next steps

The achievements of PCAWG in relation to the sharing and handling of genomic data augur well for the development of an international code that researchers everywhere can refer to.

Genomic research consortia, public and private funding bodies, and those working on existing regional codes (such as the one in Europe) might begin the process of building it. A first step would be to convene a meeting to determine the topics the code would touch on, the best way to consult research participants about their needs and a decision-making process that will allow the text to be finalized in a timely way.

If genomics researchers are instead left in the dark about how to properly address data protection and sharing, they could either be excessively cautious and fail to share as consents allow, or fail to provide participants with appropriate protection<sup>15</sup>. In other words, further regulatory uncertainty risks stalling new genomic analyses and undermining people's faith in scientific collaboration for the public good.

## The authors

**Mark Phillips** is an academic associate at the Centre of Genomics and Policy, McGill University, Montreal, Canada, and a lawyer who advises clients on data protection.

**Fruzsina Molnár-Gábor** is research group leader at the Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany.

**Jan O. Korbel** is a senior scientist at the European Molecular Biology Laboratory, Heidelberg, Germany. **Adrian Thorogood** is an academic associate at the Centre of Genomics and Policy, McGill University, Montreal, Canada. **Yann Joly** is associate professor at the Centre of Genomics and Policy, McGill University, and was the Data Access Control Officer for the International Cancer Genome Consortium (2009–18), including the PCAWG project. **Don Chalmers** is distinguished professor of law at the University of Tasmania, Hobart, Australia.

**David Townend** is professor of health and life science jurisprudence, CAPHRI Research School, Maastricht University, the Netherlands.

**Bartha M. Knoppers** is director of the Centre of Genomics and Policy, McGill University, Montreal, Canada. A full list of PCAWG Consortium Members accompanies this Comment online (see [go.nature.com/2usqhj7](https://go.nature.com/2usqhj7)). e-mails: [mark.phillips2@mcgill.ca](mailto:mark.phillips2@mcgill.ca); [fruzsina.molnar-gabor@adw.uni-heidelberg.de](mailto:fruzsina.molnar-gabor@adw.uni-heidelberg.de)

1. Rozenblatt-Rosen, O. *et al. Nature* **550**, 451–453 (2017).
2. *Nature* **523**, 136–137 (2015).
3. Stein, L. D., Knoppers, B. M., Campbell, P., Getz, G. & Korbel, J. O. *Nature* **523**, 149–151 (2015).
4. Mochama, V. 'DNA testing to aid deportations leaves plenty of room for misinterpretation and mistreatment' (*The Star*, 29 July 2018).
5. Ohm, P. *UCLA Law Rev.* **57**, 1701–1777 (2010).
6. El Emam, K. & Arbuckle, L. *Anonymizing Health Data* (O'Reilly, 2013).
7. Phillips, M. & Knoppers, B. M. *Nature Biotechnol.* **34**, 1102–1103 (2016).
8. Phillips, M., Dove, E. S. & Knoppers, B. M. *J. Bioeth. Inq.* **14**, 527–539 (2017).
9. Rocher, L., Hendrickx, J. M. & de Montjoye, Y.-A. *Nature Commun.* **10**, 3069 (2019).
10. Shringapure, S. S. & Bustamante, C. D. *Am. J. Hum. Genet.* **97**, 631–646 (2015).
11. Contreras, J. L. & Knoppers, B. M. *Annu. Rev. Genom. Hum. Genet.* **19**, 429–453 (2018).
12. Thorogood, A. *et al. Hum. Genomics* **12**, 7 (2018).
13. Thorogood, A., Dalpé, G. & Knoppers, B. M. *Eur. J. Hum. Genet.* **27**, 535–546 (2019).
14. Wolf, L. E. *et al. J. Law Med. Ethics* **43**, 594–609 (2015).
15. Knoppers, B. *Nature* **558**, 189 (2018).
16. Birney, E., Vamathevan, J. & Goodhand, P. Preprint at bioRxiv <https://doi.org/10.1101/203554> (2017).



RUTH FREEMAN/NT/REDUX/EYEVINE

A migrant farm worker has her fingerprints scanned so that she can register for a national identity card in India.

# The long road to fairer algorithms

Matt J. Kusner & Joshua R. Loftus

## Build models that identify and mitigate the causes of discrimination.

**A**n algorithm deployed across the United States is now known to underestimate the health needs of black patients<sup>1</sup>. The algorithm uses health-care costs as a proxy for health needs. But black patients' health-care costs have historically been lower because systemic racism has impeded their access to treatment – not because they are healthier.

This example illustrates how machine learning and artificial intelligence can maintain and amplify inequity. Most algorithms exploit crude correlations in data. Yet these correlations are often by-products of more salient social relationships (in the health-care example, treatment that is inaccessible is, by definition, cheaper), or chance occurrences that will not replicate.

To identify and mitigate discriminatory

relationships in data, we need models that capture or account for the causal pathways that give rise to them. Here we outline what is required to build models that would allow us to explore ethical issues underlying seemingly objective analyses. Only by unearthing the true causes of discrimination can we build algorithms that correct for these.

### Causal models

Models that account for causal pathways have three advantages. These 'causal models' are: tailored to the data at hand; allow us to account for quantities that aren't observed; and address shortcomings in current concepts of fairness (see 'Fairness four ways').

A causal model<sup>2</sup> represents how data are generated, and how variables might change in response to interventions. This can be shown as a graph in which each variable is a node and arrows represent the causal connections between them. Take, for example, a data set about who gets a visa to work in a country. There is information about the country each person comes from, the work they do, their religion and whether or not they obtained a

visa (see 'Three causal tests', part 1).

This model says that the country of origin directly influences a person's religion and whether they obtain a visa; so, too, do religion and type of work. Having a causal model allows us to address questions related to ethics, such as does religion influence the visa process?

But because many different causal models could have led to a particular observed data set, it is not generally possible to identify the right causal model from that data set alone<sup>3</sup>. For example, without any extra assumptions, data generated from the causal graph described here could seem identical to those from a graph in which religion is no longer linked to visa granting. A modeller must therefore also leverage experiments and expert knowledge, and probe assumptions.

Experiments can help in identifying factors that affect fairness. For example, a modeller wishing to explore whether ethnicity would affect treatment recommendations made online by health-care professionals could create two patient profiles that differ only in some respect that relates to ethnicity. For instance, one profile could have a name common to Americans of Chinese descent, and the other a name common to Americans of African descent. If the treatment recommendations are the same, then names can be ruled out as a source of bias, and the model can be stress-tested in another way.

Few aspects of a deep, multifaceted concept can be tested as easily as changing a name. This means that experimental evidence can underestimate the effects of discrimination. Integration of expert knowledge, particularly



from the social sciences and including qualitative methods, can help to overcome such limitations. This knowledge can be used to, for example, inform the modeller of variables that might be influential but unobserved (lighter circles in ‘Three causal tests’), or to determine where to put arrows.

Assumptions about unobserved variables that might alter the predictions of a model need to be clearly stated. This is particularly important when experiments cannot be run or more detailed expert knowledge is not available. For example, if ‘health-care access’ is not observed in a model attempting to predict ‘health need’, then it is crucial to identify any potential impacts it might have on ‘health costs’ as well as how it is affected by ‘ethnicity’.

This need for context and metadata makes causal models harder to build than non-causal ones. It can also make them a more powerful way to explore ethical questions.

### Three tests

Causal models can test the fairness of predictive algorithms in three ways.

**Counterfactuals.** A causal model allows us to ask and answer questions such as ‘Had the past been different, would the present or future have changed?’ In the visa example (see ‘Three causal tests’, part 1), algorithmic biases could be smoked out by tweaking parts of the model to explore, for instance: ‘Had individual X been Christian, would this algorithm have granted them a visa?’ A researcher could then identify what pieces of information an algorithm could use to achieve counterfactual fairness<sup>4</sup>: the algorithm’s output would not change regardless of the individual’s religion. For example, if the algorithm used just work and not country of origin or religion, it would satisfy counterfactual fairness.

**Sensitivity.** In many settings, unknowns alter knowns – data we can observe are influenced by data we cannot. Consider a causal model for a trial setting (see ‘Three causal tests’, part 2).

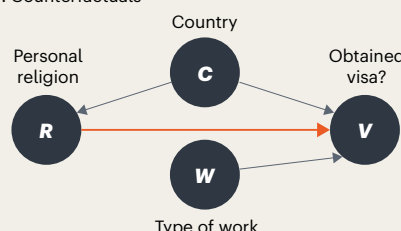
This model shows how two independent sets of unobserved quantities, structural racism and jury racism, can unfairly lead to a guilty verdict. Although researchers often cannot precisely identify unobserved variables, they can reason about how sensitive a model is to them. For instance, they can explore how sensitive our estimate of the causal link between legal representation and guilty verdict is to different levels of jury racism. Simulations of the worst-case bias scenarios (that is, when jury racism is highest) can then be used to alter jury selection to minimize the bias.

**Impacts.** Data-driven decisions can have long-term consequences and spillover effects. These effects might not be obvious, especially in the standard machine-learning paradigm

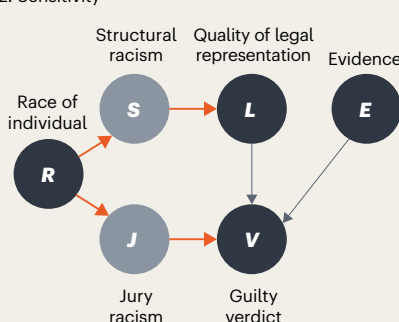
## THREE CAUSAL TESTS

Algorithmic fairness can be examined in different ways.

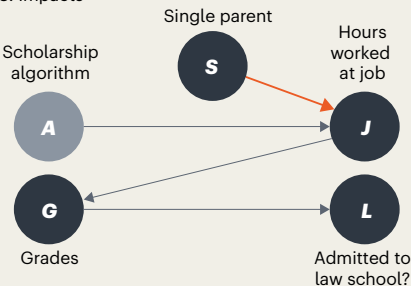
### 1. Counterfactuals



### 2. Sensitivity



### 3. Impacts



of predicting one short-term outcome. But carefully designed causal models can help researchers to use ‘interventions’ to probe the ripple effects of decisions far into the future<sup>5,6</sup>. For instance, the models can help regulatory agencies to understand how changing a scholarship algorithm influences who is accepted into law school (see ‘Three causal tests’, part 3). In this example, a single parent might need a scholarship so that they can reduce the hours they need to spend at a job, leaving them more time for study. That boosts their grades and therefore influences their chances of being admitted to law school. This complex chain can be explored using causal models.

### Five steps

Causal models are powerful tools, but they must be used appropriately. They are only models, and will thus fail to capture important aspects of the real world. Here we offer some guidelines on using them wisely.

**Collaborate across fields.** Researchers in statistics and machine learning need to know more about the causes of unfairness in society. They should work closely with those in disciplines such as law, social sciences and the humanities. This will help them to incorporate the context of the data used to train the algorithms. For example, scholars should meet at interdisciplinary workshops and conferences. One such is next year’s Association for Computing and Machinery (ACM) conference on Fairness, Accountability and Transparency to derive a set of causal models for setting bail price and for immigration decisions.

A great example of such collaborations is one between information scientist Solon Barocas at Cornell University in Ithaca, New York, and attorney Andrew Selbst at the Data & Society Research Institute in New York City. They described how current law is unable to deal with algorithmic bias<sup>7</sup>. Partly in response to this work, machine-learning researchers have launched a large subfield, known as algorithmic fairness, that looks into ways of removing bias from data. And we and other researchers now use causal models to quantify discrimination due to data.

**Partner with stakeholders.** Predictive algorithms should be developed with people they are likely to affect. Stakeholders are best placed to provide input on difficult ethical questions and historical context. One example is the work by statistician Kristian Lum at the Human Rights Data Analysis Group in San Francisco, California, which investigates criminal-justice algorithms<sup>8</sup>. Such algorithms decide whether to detain or release arrested individuals and how high to set their bail, yet they are known to be biased. Lum has invited people affected by such decisions to speak at academic conferences attended by people who research these algorithms. This has led to closer collaboration, including the tutorial ‘Understanding the context and consequences of pre-trial detention’ presented at the 2018 ACM conference on Fairness, Accountability and Transparency in New York. So far, most stakeholder work has focused on criminal justice. Another setting that would benefit from it is mortgage lending.

We propose that a rotating, interdisciplinary panel of stakeholders investigates the impacts of algorithmic decisions, for example as part of a new international regulatory institute.

**Make the workforce equitable.** Women and people from minority groups are under-represented in the fields of statistics and machine learning. This directly contributes to the creation of unfair algorithms. For example, if facial detection software struggles to detect faces of black people<sup>9</sup>, it is likely the algorithm was trained largely on data representing white people. Initiatives such as Black in AI (go.nature.com/38pbcaa) or Women in Machine Learning

## Fairness four ways

**A flurry of work has conceptualized fairness. Here are some of the most popular, and ways in which causal models offer alternatives.**

**Fairness through unawareness**<sup>12</sup>. This method works by removing any data that are considered *prima facie* to be unfair. For example, for an algorithm used by judges making parole decisions, fairness through unawareness could dictate that data on ethnic origin should be removed when training this algorithm, whereas data on the number of previous offences can be used. But most data are biased. For instance, number of previous offences can bear the stamp of historical racial bias in policing, as can the use of plea bargaining (pleading guilty being more likely to reduce a sentence than arguing innocence)<sup>13</sup>. This can leave researchers with a hard choice: either remove all data or keep biased data.

Alternatively, causal models can directly quantify how data are biased.

**Demographic parity**<sup>14</sup>. A predictive algorithm satisfies demographic parity if, on average, it gives the same predictions to different groups. For example, a university-admissions algorithm would satisfy demographic parity for gender if 50% of its offers went to women and 50% to men. It is currently more common in law to relax demographic parity so that predictions aren't necessarily equal, but are not too imbalanced. Specifically, the US Equal Employment Opportunity Commission states that fair employment should satisfy the 80% rule: the acceptance rate for any group should be no less than 80% of that of the highest-accepted group. For instance, if 25% of women were offered jobs, and this is the highest acceptance rate, then at least 20% of men must be offered

jobs<sup>4</sup>. One criticism of demographic parity is that it might not make sense to use it in certain settings, such as a fair arrest rate for violent crimes (men are significantly more likely to commit acts of violence)<sup>15</sup>.

Instead, one could require that counterfactual versions of the same individual should get the same prediction<sup>4</sup>.

**Equality of opportunity**<sup>16</sup>. This is the principle of giving the same beneficial predictions to individuals in each group. Consider a predictive algorithm that grants loans only to individuals who have paid back previous loans. It satisfies 'disability-based equality of opportunity' if it grants loans to the same percentage of individuals who both pay back and have a disability as it does to those who pay back and who do not have a disability. However, being able to pay back a loan in the first place can be affected by bias: discriminatory employers might be less likely to hire a person with a disability, which can make it harder for that person to pay back a loan. This societal unfairness is not captured by equality of opportunity.

A causal model could be used to quantify the bias and estimate an unbiased version of loan repayment.

**Individual fairness**<sup>17</sup>. This concept states that similar individuals should get similar predictions. If two people are alike except for their sexual orientation, say, an algorithm that displays job advertisements should display the same jobs to both. The main issue with this concept is how to define similar. In this example, training data will probably have been distorted by the fact that one in five individuals from sexual or gender minorities report discrimination against them in hiring, promotions and pay<sup>18</sup>. Thus similarity is hard to define, which makes individual fairness hard to use in practice.

In causal modelling, counterfactuals offer a natural way to define a similar individual. **M.J.K. & J.R.L.**

when previous attempts to address a bias failed because people strategically changed behaviours in response. In these cases, an algorithmic solution would paper over a system that needs fundamental change.

**Foment criticism.** A vibrant culture of feedback is essential. Researchers need to continually question their models, evaluation techniques and assumptions. Useful as causal models are, they should be scrutinized intensely: bad models can make discrimination worse<sup>11</sup>. At the very least, a scientist should check whether a model has the right data to make causal claims, and how much these claims would change when the assumptions are relaxed.

Algorithms are increasingly used to make potentially life-changing decisions about people. By using causal models to formalize our understanding of discrimination, we must build these algorithms to respect the ethical standards required of human decision makers.

## The authors

**Matt J. Kusner** is an associate professor in the Department of Computer Science at University College London, UK. **Joshua R. Loftus** is an assistant professor in the Department of Technology, Operations, and Statistics at New York University, New York, USA. e-mails: matt.kusner@gmail.com; loftus@nyu.edu

1. Obermeyer, Z. *et al.* *Science* **366**, 447–453 (2019).
2. Judea, P. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2000).
3. Spirtes, P. *et al.* *Causation, Prediction, and Search* (MIT Press, 2000).
4. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. In *Advances in Neural Information Processing Systems* 4066–4076 (MIT Press, 2017).
5. Liu, L. T. *et al.* In *International Conference on Machine Learning* 3150–3158 (ACM, 2018).
6. Kusner, M., Russell, C., Loftus, J. & Silva, R. *Proc. Machine Learning Res.* **97**, 3591–3600 (2019).
7. Barocas, S. & Selbst, A. D. *Calif. L. Rev.* **104**, 671 (2016).
8. Lum, K. *Nature Hum. Behav.* **1**, 0141 (2017).
9. Simon, M. 'HP looking into claim webcams can't see black people.' (CNN Tech, 23 December 2009).
10. McManus, H. D. *et al.* *Race Justice* <https://doi.org/10.1177/2153368719849486> (2019).
11. Kilbertus, N. *et al.* 'The Sensitivity of Counterfactual Fairness to Unmeasured Confounding'. In *Uncertainty in Artificial Intelligence* (AUAI, 2019).
12. Grigic-Hlaca, N. *et al.* 'The case for process fairness in learning: Feature selection for fair decision making.' *NeurIPS Symposium on Machine Learning and the Law* (2016).
13. Wilford, M. M. & Khairalla, A. In *Social Sciences Contributions to the Real Legal System* Ch. 7, 132 (2019).
14. Zafar, M. B., Valera, I., Rodriguez, M. G. & Gummadi, K. P. In *Artificial Intelligence and Statistics* 962–970 (2017).
15. Dobash, R. E., Dobash, R. P., Cavanagh, K. & Lewis, R. *Violence Against Women* **10**, 577–605 (2004).
16. Hardt, M., Price, E. & Srebro, N. 'Equality of opportunity in supervised learning'. In *Advances in Neural Information Processing Systems* 3315–3323 (2016).
17. Dwork, C. *et al.* 'Fairness through awareness'. In *Proc. 3rd Innov. Theoret. Comp. Sci. Conf.* 214–226 (2012).
18. Pizer, J. C. *et al.* *Loy. LAL Rev.* **45**, 715 (2011).

(go.nature.com/2s5km5g) are positive steps.

And we can go further. Causal models can themselves help to address the field's 'pipeline problem' by identifying where unfairness enters the process and which interventions can increase the participation of under-represented groups without shifting the burden to extra work for role models in those groups. Academic institutions should critically evaluate and use these models for fairer admissions in fields related to artificial intelligence.

**Identify when algorithms are inappropriate.** Statistics and machine learning are not all-powerful. Some problems should

not be solved by expanding data-gathering capabilities and automating decisions. For example, a more accurate model for predictive policing won't solve many of the ethical concerns related to the criminal legal system. In fact, these methods can mask structural issues, including the fact that many neighbourhoods are policed by people who do not live in them<sup>10</sup>. This disconnect means that police officers might not be invested in the community they police or the people they arrest.

There are red flags when demographics, such as ethnic origin, influence nearly every piece of information in a causal graph, or



# Correspondence

## Ongoing horror of 2019 oil disaster

The mystery crude-oil spill that struck Brazil from late August last year continues to severely affect thousands of kilometres of the country's northeastern coastline. Remediation and containment measures are being hampered because the source and timing of the spill are still unclear. The consequences of this environmental and societal disaster could last for decades.

Besides the spill's catastrophic impact on the region's marine biodiversity (more than 40 of Brazil's Marine Protected Areas have been hit), it affects the livelihoods and food security of millions of coastal residents. In the region of Pernambuco, for example, sales of fish and shellfish have plummeted by around 80% (M. E. de Araújo *et al. Cad. Saúde Públ.* <http://doi.org/dkq7>; 2020).

The decline in sales of fish and seafood in the region is exacerbated by public confusion over the authorities' conflicting advice on safe eating habits. The biological accumulation of toxins in food animals is likely to pose a long-term risk to human health.

**Richard J. Ladle\*** Institute of Biological and Health Sciences, Federal University of Alagoas, Maceió, Brazil.

\*On behalf of 4 correspondents; see [go.nature.com/2tz2u13](https://go.nature.com/2tz2u13)  
richardjamesladle@gmail.com

## Don't cheat Chinese environment laws

Some local authorities in China are resorting to 'quick fixes' to comply with strict regulations imposed by the 2014 revised Environmental Protection Law. Such tactics must be stopped: they mask pollution issues that could defeat long-term environmental goals.

For example, to pre-empt scrutiny and quickly improve air-quality rankings in Lanshan, Shandong province, 300 or so restaurants were shut and production at more than 400 wooden-fibreboard factories was stopped. Many of these businesses had pollution-control measures in place ([www.mee.gov.cn](http://www.mee.gov.cn)). In another case, farmers in Shangcai, Henan, were told to harvest 5 hectares of wheat by hand to avoid dust from mechanized harvesting affecting readings at a nearby air-monitoring site (J. Qu *Chutian Metropolis Daily* 10 June 2019). Air-quality monitoring data have also been manipulated in Xi'an, Shaanxi (D. Liu and S. Wang *Int. J. Environ. Sci. Technol.* **16**, 4963–4966; 2019).

Local leaders' success should be measured by their progress on long-term environmental improvements. Regional governments should therefore be allowed a reasonable period to address environmental issues. The focus should be on tackling underlying causes of environmental problems, with technical backing from central government, rather than on misguided quick fixes.

**Dasheng Liu** Ecological Society of Shandong, Jinan, China.  
ecologyliu@163.com

**Renqing Wang** Shandong University, Qingdao, China.

**Julian R. Thompson** University College London, UK.

## Ditch group metrics for student hopefuls

As an associate professor at an Australian university who was educated at unranked universities in India, I find it disturbing that some universities are now using international university rankings to help assess graduate students for admission. In my view, this risks promoting and institutionalizing discrimination, and hence undermines global efforts to increase diversity in academia.

When I applied in 1998 to do a PhD at the University of Zurich, Switzerland, the university requested my degree-course syllabuses from India. My opportunities were not scuttled by the ranking of those universities. So I was shocked when one of my students showed me the applications section for master's programmes at several premier institutions. These required applicants to give the ranking of the university where they studied as an undergraduate, for use as an assessment parameter.

Such 'objective' metrics could be viewed as a way to reduce the selection workload and avoid unconscious biases. But individuals should not be assessed through a group-based metric that reinforces stereotypes. And, given that university rankings are correlated with per capita gross domestic product (E. F. Tuesta *et al. J. Data Inf. Sci.* **4**, 56–78; 2019), organizations also risk making the serious mistake of equating an applicant's ability with regional and economic differences.

**Sureshkumar Balasubramanian** Monash University, Melbourne, Australia.  
mb.suresh@monash.edu

## No foundation for anti-nuclear bias

In his otherwise excellent review of Thane Gustafson's book *The Bridge*, Andrew Moravcsik includes nuclear power in his list of energy sources to which natural gas is "environmentally superior" (*Nature* **576**, 30–31; 2019). Burning natural gas in fact releases almost half as much carbon dioxide into the atmosphere as does coal. Nuclear power production itself releases none. Taking into account the CO<sub>2</sub> released from fossil fuels burnt during plant construction and uranium mining and processing, nuclear energy ranks about equally with solar power – so still much less polluting than natural gas.

If Moravcsik is referring to damage from environmental releases, nuclear power has proved itself to be much cleaner and safer than natural gas. If he is considering nuclear waste, existing practice effectively sequesters spent nuclear fuel from the environment by using dry-cask storage. Permanent disposal sites (two currently: the Waste Isolation Pilot Plant in southern New Mexico, for military waste, and Onkalo, under construction in Finland) will effectively isolate nuclear waste for centuries.

Anti-nuclear bias has no place in a pre-eminent journal of science, especially when global heating is increasing dangerously as a result of our over-dependence on fossil fuels.

**Richard Rhodes**, Half Moon Bay, California, USA.  
richardrhodes1@comcast.net

# News & views

## Tumour genetics

# Global cancer genomics project comes to fruition

Marcin Cieslik & Arul M. Chinnaiyan

A massive international effort has yielded multifaceted studies of more than 2,600 tumours from 38 tissues, generating a wealth of insights into the genetic basis of cancer. See p.82, p.94, p.102, p.112, p.122 & p.129

Comprehensive genomic characterization of tumours became a major goal of cancer researchers as soon as the first human genome had been sequenced in 2001. Since then, advances in sequencing technology and analytical tools have allowed this research field to flourish. In six papers<sup>1–6</sup> in this issue of *Nature*, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium presents the most comprehensive and ambitious meta-analysis of cancer genomes so far. Unlike previous efforts that focused largely on protein-coding regions of the cancer genome, PCAWG analyses whole genomes. Each article scrutinizes an important aspect of cancer genetics – together, their findings will be key to understanding the full genetic complexity of cancer.

Before discussing the impact of these analyses, it is crucial to highlight the massive amount of data and the complex organizational framework that underpin the PCAWG endeavour. The project involved an interdisciplinary group of scientists from 4 continents, with 744 affiliations between them, who had to overcome major technical, legal and ethical challenges to carry out distributed analyses while protecting patient data. Researchers were divided into 16 working groups, each focused on distinct facets of cancer genomics – assessing the recurrence of mutations, for instance, or inferring tumour evolution.

Altogether, the consortium performed integrative analyses of 38 tumour types. The group sequenced 2,658 whole-cancer genomes (Fig. 1), alongside matched samples of non-cancerous cells from the same individuals. These data were complemented by 1,188 transcriptomes – the sequences and abundances of RNA transcripts in a tumour.

These efforts involved extensive quality control and coordinated data processing,

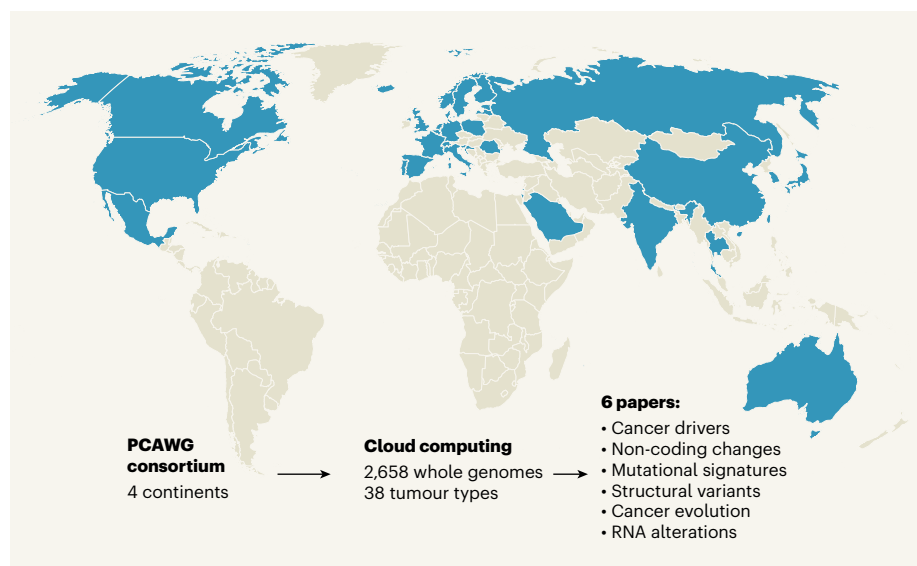
as well as massive, systematic experimental validation of the computational pipelines used to detect mutations. Many computational algorithms and pipelines were used and compared in concert. This required hundreds of terabytes of data, spread across multiple data centres, and probably millions of processing hours – all facilitated by cloud computing. Notably, the PCAWG efforts provide a prime example of how cloud computing can make international collaboration possible and help to advance data-intensive fields.

The first of the current papers<sup>1</sup> (page 82) gives an overview of the breadth and depth of

the PCAWG data set. The consortium reports that, on average, each cancer genome carries four or five driver mutations, which provide cancer cells with a selective advantage. Only 5% of tumours studied had no identified driver aberrations. By contrast, many cancers exhibited hallmarks of genomic catastrophes called chromoplexy (17.8% of tumours) and chromothripsis (22.3%), which result in major structural changes to the genome.

The other five papers each delve into a different aspect of the data set in more detail. For instance, in the second paper, Rheinbay *et al.*<sup>2</sup> (page 102) set out to identify genetic drivers in non-coding DNA. This is an ambitious undertaking, because it is substantially more difficult to accurately detect mutations in non-coding regions than in coding regions, or to assess their recurrence. The authors used careful modelling to exclude artefacts and systematically identify non-coding driver mutations.

Their results call into question previously reported non-coding drivers, such as the long non-coding RNAs *NEAT1* and *MALAT1*, but also reveal new ones. For example, the authors report a recurrent mutation in a non-coding region of the key tumour-suppressor gene *TP53*. They also found relatively frequent mutations in non-coding regions of the telomerase gene *TERT* that result in over-expression of the telomerase enzyme (which



**Figure 1 | A worldwide effort to tackle cancer.** The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium is a group of cancer researchers from four continents (blue). The group sequenced and analysed 2,658 whole-cancer genomes from 38 types of tumour. The huge amount of data involved in the effort required sophisticated cloud-computing approaches. Six papers<sup>1–6</sup> from the PCAWG now describe different aspects of the analyses performed. (Nature publications remain neutral with regard to contested jurisdictional claims in published maps.)



helps tumour cells to divide uncontrollably), mirroring the high (12%) prevalence of telomerase mutations found in a previous pan-cancer study of more-advanced (metastatic) tumours<sup>7</sup>. Although the study could not rule out the existence of other non-coding drivers, it decisively shows that this type of mutation is not common.

In the third and fourth papers, Alexandrov *et al.*<sup>3</sup> (page 94) and Li *et al.*<sup>4</sup> (page 112) focus on genomic aberrations called signatures. Different processes, such as defective DNA-repair mechanisms or exposure to environmental mutagens, produce these characteristic patterns of DNA aberrations. Large genomic data sets are crucial if we are to refine known mutational signatures and discover new ones. Impressively, between them, Alexandrov *et al.* and Li *et al.* identify 97 signatures. This expansion on previous work encompasses not only conventional single-nucleotide signatures, but also signatures involving multi-nucleotide variants and small insertions or deletions of DNA.

Notably, Li and colleagues are among the first to uncover reproducible signatures involving structural variants (SVs) – rearrangements of large portions of the genome. The process was much more intricate than that for identifying mutational signatures because of the diversity and complexity of SVs.

Through a series of mutation-subgrouping steps, the researchers identified 16 SV signatures, revealing, for example, a putative mechanistic link between two SVs, deletions and reciprocal inversions (the last of which involves a reversal of the orientation of a segment of DNA). They also gained insights into the roles of all 16 signatures in cancer. Mutations in certain DNA-repair genes were shown to associate with characteristic cancer signatures. For instance, the consortium found that mutations in the gene *CDK12* associate with tandem stretches of duplicated DNA, and that truncated variants of the DNA-repair enzyme MBD4 co-occur with a distinct mutational signature involving DNA sequences called CpG sites. Altogether, these new signatures lay the foundation for understanding mechanisms of cancer development, and the role of mutagenic exposures in this process.

The idea that cancer develops through an evolutionary process was first presented in 1976 (ref. 8). Since then, cancer evolution has been characterized in terms of random mutations and natural selection. A cancer cell harbouring a mutation that confers high fitness proliferates rapidly, becoming the most prominent cell clone in the population. This phenomenon, called a clonal sweep, occurs in recurring cycles during cancer growth. Cancer evolution is most effectively studied by sequencing multiple regions of a tumour over time, but it can also be reconstructed from a single biopsy – the approach taken by Gerstung *et al.*<sup>5</sup> (page 122) in the fifth paper.

The authors introduce the concept of ‘molecular time’ to classify clonal and subclonal mutations. They reasoned that subclonal mutations, which are present in only a subset of a tumour’s cells, must have arisen late in the cancer’s evolution. They classify clonal mutations, which are present in all of a tumour’s cells, as early or late, depending on whether the mutations arose before or after the clone underwent copy-number gains (an increase in the number of copies of a gene or chromosomal region). The researchers aggregated evolutionary data from multiple tumours, allowing them to identify common mutational trajectories such as *APC*–*KRAS*–*TP53* progression<sup>9</sup>, which describes the typical sequence in which mutations arise in colorectal cancer.

Gerstung *et al.* found that the driver mutations that most commonly occur in a given cancer also tend to occur the earliest. Similarly, if copy-number gains are highly recurrent in a particular cancer type, they tend to occur early. For example, a copy-number gain in part of chromosome 5 is common in clear cell kidney cancer, and tends to arise early in the disease’s evolution. Conversely, whole-genome duplication is a relatively late event in this cancer. Finally, the researchers found that mutational

### “The broad availability and quality of the new data set will almost certainly spur a wave of biological insights.”

signatures change over time in at least 40% of tumours. These changes reflect a decreasing role for environmental exposures in disease progression and an increase in the frequency and severity of DNA-repair defects. Overall, the group’s findings suggest that driver mutations can occur years before cancer is diagnosed, which has implications for early detection and biomarker development.

In the final paper (page 129), the PCAWG Transcriptome Core Group and their colleagues<sup>6</sup> made use of the 1,188 PCAWG samples that had matched transcriptome data, to functionally link DNA and RNA alterations. The group found associations between hundreds of single-nucleotide DNA mutations and the expression of nearby genes. However, larger copy-number alterations were the main drivers of gene-expression changes in cancer cells. Mutations were also associated with changes in transcript structure, such as the formation of a new protein-coding region (an exon) within a non-coding region (an intron).

The authors also characterized the frequency of bridged fusions – a phenomenon in which two genes become fused owing to a third, intervening fragment of DNA. Finally, although 87 of the 1,188 samples analysed did not have a driver alteration at the DNA level, the

group showed that every one of these had an RNA-level alteration. Together, these insights illustrate the power of integrated RNA- and DNA-sequencing analysis for cancer studies<sup>10</sup>.

These six papers, together with companion papers being co-published elsewhere (see [go.nature.com/3boajsm](http://go.nature.com/3boajsm)), represent a milestone in cancer and cloud genomics. By focusing on inferences, the PCAWG successfully expands on a decade of cancer sequencing studies that were rooted largely in observations. It is worth noting that, although inferential analyses offer a deeper look at cancer than do descriptive studies, their results are also associated with a higher degree of uncertainty.

The broad availability and quality of the PCAWG data set will almost certainly spur a wave of biological insights and methodological developments. Integration with other functional genomic data sets, for example probing the 3D organization of the genome, will also undoubtedly provide further understanding of the causes and consequences of genetic aberrations.

The biggest limitation of the current studies is the lack of clinical data concerning patient outcomes and treatments. Such data would allow researchers to identify the genetic changes that can predict clinical outcomes. Fortunately, a project called the International Cancer Genome Consortium–Accelerate Research in Genomic Oncology (ICGC–ARGO) is under way that will create such a resource for more than 100,000 people with cancer.

Ultimately, the PCAWG brought together thousands of scientists, working together to achieve its aims. The long-term impact of these efforts will not be limited to the findings published today, but will also come from the collaborations that have formed and the knowledge exchanges that have taken place between members of this global consortium of researchers.

**Marcin Cieslik** and **Arul M. Chinnaiyan** are in the Michigan Center for Translational Pathology, Rogel Cancer Center, University of Michigan, Ann Arbor, Michigan 48109, USA. **M.C.** is also in the Department of Computational Medicine and Bioinformatics, University of Michigan. **A.M.C.** is also at the Howard Hughes Medical Institute. e-mails: [mcieslik@med.umich.edu](mailto:mcieslik@med.umich.edu); [arul@med.umich.edu](mailto:arul@med.umich.edu)

1. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. *Nature* **578**, 82–93 (2020).
2. Rheinbay, E. *et al.* *Nature* **578**, 102–111 (2020).
3. Alexandrov, L. B. *et al.* *Nature* **578**, 94–101 (2020).
4. Li, Y. *et al.* *Nature* **578**, 112–121 (2020).
5. Gerstung, M. *et al.* *Nature* **578**, 122–128 (2020).
6. PCAWG Transcriptome Core Group *et al.* *Nature* **578**, 129–136 (2020).
7. Priestley, P. *et al.* *Nature* **575**, 210–216 (2019).
8. Nowell, P. C. *Science* **194**, 23–28 (1976).
9. Fearon, E. R. & Vogelstein, B. *Cell* **61**, 759–767 (1990).
10. Robinson, D. R. *et al.* *Nature* **548**, 297–303 (2017).

# A platform for making and transferring oxide films

Atsushi Tsukazaki

Crystalline films of technologically useful oxide materials have been grown by a method based on surface-modified substrates. Unlike usual oxide films, these can be easily transferred to any material. **See p.75**

Inorganic compounds that contain oxygen and at least two other elements are known as complex oxides. Crystalline films of these compounds have desirable properties such as superconductivity, magnetism and ferroelectricity (spontaneous electric polarization), and could be used in next-generation devices<sup>1–3</sup> if they can be integrated with mature device technologies. Integration is typically achieved by growing films on compatible substrates using a method called epitaxy, but this approach works only for relatively limited material systems. In the past few years, free-standing membranes of certain oxides have been made<sup>4,5</sup> by removing epitaxial films from substrates using a process dubbed chemical lift-off. Now, on page 75, Kum *et al.*<sup>6</sup> report a versatile method for producing a wide variety of complex-oxide films that can be easily transferred to any material.

The authors' approach uses a technique known as remote epitaxy<sup>7–9</sup>, in which the epitaxial film and the substrate are separated by

a few sheets of the two-dimensional material graphene (Fig. 1a). Potential-energy fields produced by atoms in the substrate can penetrate the graphene and transmit information about the substrate's crystal lattice, enabling the epitaxial growth of high-quality films. The field penetrability is proportional to the strength of ionic bonds in the substrate material<sup>8</sup>. A film grown in this way can be easily removed (exfoliated) from the graphene because the two materials are coupled by only weak van der Waals forces (Fig. 1b). Remote epitaxy therefore combines outstanding epitaxial growth and exfoliation.

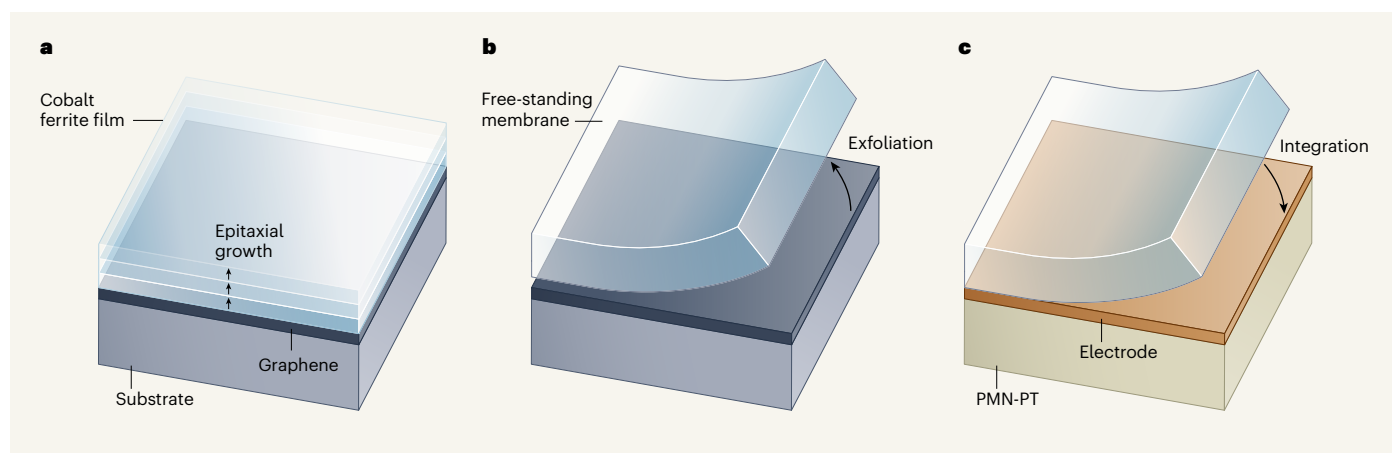
The fabrication of high-quality oxide films requires a well-regulated growth scheme and atomic-level control over the material interfaces and substrate surfaces<sup>1–3</sup>. In the past few decades, single-crystal oxide substrates of various crystal structures have become commercially available. These include substrates of strontium titanate, aluminium oxide and magnesium aluminate, which have perovskite,

corundum and spinel crystal structures, respectively.

For the epitaxial growth of a particular film, the substrate should be appropriately selected in terms of its crystal structure, lattice dimensions and coefficient of thermal expansion – a quantity that describes how the size of a material is affected by a change in temperature. Consequently, growth conditions, such as temperature, oxygen pressure and growth rate, need to be optimized to stabilize the desired crystalline phase and obtain high crystallinity.

In their remote-epitaxy work, Kum and colleagues carefully optimized the growth conditions of oxide films on graphene-coated substrates. In general, control over the degree of oxidation is crucial for making high-quality oxide films, so the background oxygen pressure should be well regulated during film growth. However, when the authors used a growth method called pulsed-laser deposition and supplied oxygen to their set-up at the required high temperature, they found that the graphene was etched from the substrate. To prevent this etching, they grew the initial part of the oxide film (a thickness of about 5–10 nanometres, compared with a final thickness of the order of 100 nm) in a vacuum. The crystallinity of this part was still high owing to oxidation of the film during the growth of the remaining part. Finally, the authors exfoliated the oxide film from the graphene to produce a free-standing oxide membrane.

In other experiments, Kum *et al.* found that strontium ruthenate could be used instead of graphene to grow an oxide film by a process known as sputtering. The film could then be exfoliated from the strontium ruthenate by depositing a layer of nickel on top of the film. The nickel acts as a stressor – it provides



**Figure 1 | Growth and integration of complex-oxide films.** **a**, Kum *et al.*<sup>6</sup> report a versatile approach for making high-quality films of technologically useful compounds called complex oxides and transferring these films to other materials. As a demonstration, the authors grew a film of the complex oxide cobalt ferrite using a technique known as remote epitaxy, whereby the film and the underlying substrate are separated by a few sheets of the material graphene. **b**, They then exfoliated (removed) the film from the graphene

to produce a free-standing membrane. **c**, Finally, the authors integrated the membrane with an electrode and a membrane of another complex oxide, lead magnesium niobate–lead titanate (PMN-PT), which had also been made using remote epitaxy (but replacing graphene with strontium ruthenate). Such integration is difficult to achieve using the conventional scheme for growing oxide films because cobalt ferrite and PMN-PT have different crystal structures.



enough strain energy to overcome the weak bonds between the film and the strontium ruthenate.

Kum and co-workers demonstrated the transferral of oxide membranes to other materials for: strontium titanate, yttrium iron garnet and magnetic cobalt ferrite, produced by pulsed-laser deposition; lead magnesium niobate–lead titanate (PMN-PT), formed by sputtering; and ferroelectric barium titanate, made by a process called molecular-beam epitaxy. One example of a stacked structure produced by such transferral consists of a 300-nm-thick layer of cobalt ferrite, an electrode and a 500-nm-thick layer of PMN-PT (Fig. 1c).

The authors found that this structure displays high magnetostriction (coupling between magnetic and mechanical behaviour) and piezoelectricity (coupling between electric and mechanical behaviour), because it is free-standing rather than being clamped by a substrate. Cobalt ferrite, PMN-PT and yttrium iron garnet have different crystal structures, making it difficult to stack these materials by the usual growth scheme without such clamping.

Kum *et al.* also stacked graphene and oxide membranes to examine the electrical coupling between these materials. The density of electric charge in graphene can be inferred from the positions of peaks in Raman spectra – spectra generated through the scattering of incident light. The authors found that these positions depend on the stacked structure, indicating that charge is transferred across graphene–membrane interfaces. These results suggest that stacks of other combinations of materials will offer ways to integrate the various functions of oxides with mature device technologies.

The authors' exfoliation technique enables complex-oxide films to be easily transferred from an epitaxial interface to any material. Because the thickness and stacking of films can be controlled, ultrathin membranes and stacks of various membranes could be possible. Such a simple way of transferring the functions of oxides might advance the field of oxide-based electronics through integration with emerging quantum material systems<sup>10</sup>. However, the availability of graphene-coated substrates could be a key issue for developing the method.

This technique will probably be extended beyond the transferral of complex-oxide films. For example, it might provide an innovative strategy for engineering interfaces, by allowing 2D or 3D films and membranes to be integrated with each other through effects associated with multiple couplings between them. An understanding of the chemical or physical bonds at the interface between membranes in stacked structures is crucial and will reveal how such an interface

differs from the epitaxial one. Finally, unusual material combinations (in which, for example, the size and orientation of crystal lattices of membranes are mismatched) could enable useful interface functions that are difficult to achieve or control at the epitaxial interface.

**Atsushi Tsukazaki** is at the Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan.  
e-mail: tsukazaki@imr.tohoku.ac.jp

### Virology

## Latent HIV-1 gets a shock

**Mathias Lichterfeld**

HIV-1 can evade the immune system by hiding out in a dormant form. Two studies describe interventions that can effectively reactivate the latent virus in animals, potentially rendering it vulnerable to immune-mediated death. **See p.154 & p.160**

'Shock and kill' might sound like a military strategy, but in fact it describes the dominant model currently used in the search for a cure for HIV-1 infection. Although antiretroviral therapy (ART) is highly effective at limiting the extent of the infection, the virus can hide out in a 'latent' form in immune cells called CD4<sup>+</sup> T cells, undergoing little or no transcription and thus remaining undetected by the immune system<sup>1,2</sup>. When ART is stopped, these viral-reservoir cells can rapidly fuel HIV rebound. The theory behind 'shock and kill'

**"The current studies showcase some of the conceptual and technical challenges intrinsically associated with pharmacological latency reversal."**

involves the use of drugs that reverse this latency and could increase viral gene expression (shock), rendering the viral-reservoir cells vulnerable to elimination (kill) by other cells of the immune system. Two groups<sup>3,4</sup> now describe distinct interventions in animal models that cause what seem to be the most robust and reproducible disruptions of viral latency reported so far.

In the first study, Nixon *et al.*<sup>3</sup> (page 160) focus on a drug called AZD5582, which can activate the transcription factor NF- $\kappa$ B – a major instigator of HIV-1-gene expression. AZD5582 was originally developed to treat

1. Ramesh, R. & Schlom, D. G. *MRS Bull.* **33**, 1006–1014 (2008).
2. Schlom, D. G., Chen, L.-Q., Pan, X., Schmehl, A. & Zurbuchen, M. A. *J. Am. Ceram. Soc.* **91**, 2429–2454 (2008).
3. Hwang, H. Y. *et al.* *Nature Mater.* **11**, 103–113 (2012).
4. Lu, D. *et al.* *Nature Mater.* **15**, 1255–1260 (2016).
5. Ji, D. *et al.* *Nature* **570**, 87–90 (2019).
6. Kum, H. S. *et al.* *Nature* **578**, 75–81 (2020).
7. Kim, Y. *et al.* *Nature* **544**, 340–343 (2017).
8. Kong, W. *et al.* *Nature Mater.* **17**, 999–1004 (2018).
9. Bae, S.-H. *et al.* *Nature Mater.* **18**, 550–560 (2019).
10. Tokura, Y., Kawasaki, M. & Nagaosa, N. *Nature Phys.* **13**, 1056–1068 (2017).

cancer, and activates the 'non-canonical' NF- $\kappa$ B pathway, which results in an atypical type of NF- $\kappa$ B-driven transcription that is slow but persistent. The authors tested AZD5582 in two animal models: 'humanized' mice (which carry human-derived liver, bone-marrow and thymus cells) that were infected with HIV; and rhesus macaques infected with the HIV-related simian immunodeficiency virus (SIV). Both groups of animals were already receiving ART.

The authors demonstrated that AZD5582 treatment led to marked increases in the levels of viral RNA in CD4<sup>+</sup> T cells in a range of tissues in both species, indicating that transcription of the virus had been activated. This was combined with a substantial rise in virus levels in the blood. AZD5582 is not optimized for use in humans; nonetheless, these results suggest that pharmacological activation of the non-canonical NF- $\kappa$ B pathway could be an attractive way to trigger HIV-1-gene expression as part of a shock-and-kill approach (Fig. 1).

In the second study, McBrien *et al.*<sup>4</sup> (page 154) used an entirely different, though complementary, approach to disrupting viral latency. Again, the authors used both ART-treated humanized mice infected with HIV-1 and ART-treated, SIV-infected rhesus macaques. They combined two immunological interventions. The first involves antibody-mediated depletion of CD8<sup>+</sup> T cells – immune cells previously shown to act in concert with ART to reduce levels of viral transcription<sup>5</sup>. The second, administered concurrently, involves treatment with a drug called N-803, which strongly activates the signalling

molecule interleukin-15 (IL-15), and which has been previously shown<sup>6</sup> to activate HIV-1 transcription *in vitro*. Like Nixon and colleagues, the researchers found that their treatment caused substantial increases in virus levels in the blood, and in viral RNA in cells from various tissues.

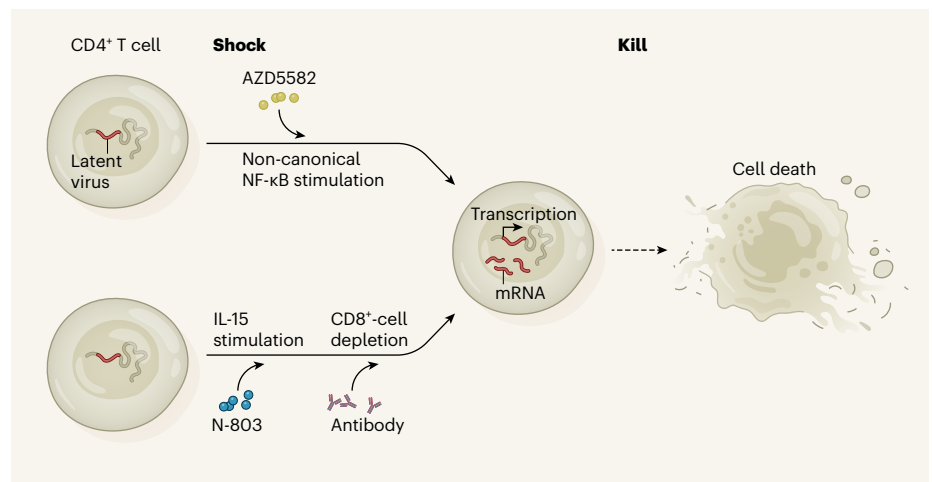
At first glance, the combined interventions used by McBrien and colleagues might seem contradictory, because IL-15 is one of the strongest activators of CD8<sup>+</sup> T cells<sup>7,8</sup>. But the synergistic effects of these two interventions raise the provocative possibility that the best strategies for targeting viral-reservoir cells involve a mix of immune interventions – suppressing immune components that seem to have a role in stabilizing viral latency (such as CD8<sup>+</sup> T cells) while activating others that can effectively disrupt latency (such as IL-15 signalling).

How exactly CD8<sup>+</sup> T-cell depletion interacts with IL-15 to reverse HIV-1 latency is unknown. Given the vast array of direct and indirect effects resulting from depletion of CD8<sup>+</sup> T cells<sup>9</sup>, it will not be easy to define the precise molecular mechanisms underlying this synergy. But an understanding of this relationship might reveal downstream proteins that are jointly targeted by these interventions and that could therefore be used to optimize latency reversal in the clinic.

In addition to the advances they make, the current studies showcase some of the conceptual and technical challenges intrinsically associated with pharmacological latency reversal. First, the latency-reversing agents (LRAs) evaluated (as well as all other LRAs described so far<sup>10</sup>) target factors that have crucial roles in modulating host-cell gene transcription, in addition to viral transcription. Their use therefore comes with an intrinsic risk of toxic off-target effects. The toxicity of the LRAs described by McBrien *et al.* and Nixon *et al.* seems to be acceptable in animal models, with most showing no clinical side effects. However, much more stringent safety standards must be met in human clinical trials.

Mechanisms of viral latency might vary between individual viral-reservoir cells and are likely to be influenced by the position at which the HIV-1 genomes have integrated into the host-cell chromosomes<sup>11</sup>. It is therefore possible that only subsets of cells will respond to individual LRAs, which typically target one specific mechanism of viral latency. The actual proportion of viral-reservoir cells that responded to the interventions in the two current studies is uncertain, and would be difficult to determine experimentally<sup>12</sup>.

Another uncertainty is how much of the increase in HIV-1 RNA is attributable to CD4<sup>+</sup> T cells carrying HIV-1 that can replicate effectively<sup>13,14</sup>. This is of interest because most viral-reservoir cells harbour HIV-1 genomes



**Figure 1 | Two approaches to reactivating dormant HIV-1.** HIV-1 can integrate into the genome of CD4<sup>+</sup> T cells in a latent form – it is not transcribed into messenger RNA and so is not detected by the body's immune system. Two papers describe 'shock' treatments that can reactivate transcription of latent HIV in mice and the related virus SIV in monkeys. Nixon *et al.*<sup>3</sup> used a drug called AZD5582 to activate the non-canonical NF-κB signalling pathway, which stimulates virus transcription. McBrien *et al.*<sup>4</sup> used two interventions – a drug called N-803 to stimulate the protein IL-15, which promotes transcription, and an antibody treatment that depletes immune cells called CD8<sup>+</sup> T cells, which seem to have a role in dampening HIV transcription. After these shock treatments have reactivated the virus, interventions that target and kill the virus-carrying CD4<sup>+</sup> T cells should help to eliminate the latent viral reservoir. Such treatments remain to be designed.

that contain lethal sequence defects, probably as a result of errors introduced during reverse transcription of viral RNA, which produces the viral DNA that is integrated into the host genome. These defective viral genomes can often still be transcribed and respond to LRAs, but they cannot cause viral rebound when ART is stopped and so do not represent the main target for shock-and-kill interventions. In addition, it is unclear how disrupting latency might influence the evolutionary dynamics of the reservoir cells – whether, for instance, a shock treatment kills some subsets of CD4<sup>+</sup> T cells that are highly susceptible to latency disruption, but confers a selective advantage on other subsets of non-susceptible, difficult-to-reactivate cells.

Most importantly, neither of the interventions tested in the current studies led to a change in the expression of markers of viral-reservoir size. A decrease in these markers is the most informative and crucial endpoint parameter for shock-and-kill approaches. The absence of an effect on viral-reservoir size probably reflects the fact that the studies were mainly designed to investigate latency reversal, and lacked dedicated 'kill' interventions. Combining 'shock' interventions with 'kill' components is a key next step. In fact, that they provide a suitable model for evaluating 'kill' strategies in the setting of robust and efficient latency reversal might be one of the strengths of the current studies.

Finally, the work of Nixon and colleagues and McBrien and colleagues should not distract from the fact that the shock-and-kill strategy currently remains largely a theoretical

concept, not a therapeutic reality. Establishing evidence for its ability to reduce viral reservoirs and to deliver real benefits to patients will require much more work.

**Mathias Lichterfeld** is in the Infectious Disease Division, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. He is also at the Ragon Institute of MGH, MIT and Harvard and at the Broad Institute of MIT and Harvard, both in Cambridge, Massachusetts, and in the Joint Center for Human Retrovirus Infection, Kumamoto University, Japan.  
e-mail: mlichterfeld@partners.org

1. Finzi, D. *et al.* *Nature Med.* **5**, 512–517 (1999).
2. Ruelas, D. S. & Greene, W. C. *Cell* **155**, 519–529 (2013).
3. Nixon, C. C. *et al.* *Nature* **578**, 160–165 (2020).
4. McBrien, J. B. *et al.* *Nature* **578**, 154–159 (2020).
5. Cartwright, E. K. *et al.* *Immunity* **45**, 656–668 (2016).
6. Jones, R. B. *et al.* *PLoS Pathog.* **12**, e1005545 (2016).
7. Conlon, K. C. *et al.* *J. Clin. Oncol.* **33**, 74–82 (2015).
8. Younes, S.-A. *et al.* *J. Clin. Invest.* **126**, 2745–2756 (2016).
9. Okoye, A. *et al.* *J. Exp. Med.* **206**, 1575–1588 (2009).
10. Spivak, A. M. & Planelles, V. *Annu. Rev. Med.* **69**, 421–436 (2018).
11. Chen, H.-C., Martinez, J. P., Zorita, E., Meyerhans, A. & Filion, G. J. *Nature Struct. Mol. Biol.* **24**, 47–54 (2017).
12. Cillo, A. R. *et al.* *Proc. Natl Acad. Sci. USA* **111**, 7078–7083 (2014).
13. Ho, Y.-C. *et al.* *Cell* **155**, 540–551 (2013).
14. Lee, G. Q. *et al.* *J. Clin. Invest.* **127**, 2689–2696 (2017).

This article was published online on 22 January 2020.



## Accelerator physics

# Muon colliders come a step closer

Robert D. Ryne

Particle colliders that use elementary particles called muons could outperform conventional colliders, while requiring much smaller facilities. Muon cooling, a milestone on the road to these muon colliders, has now been achieved. **See p.53**

"SMASH! Colossal colliders are unlocking the secrets of the universe." The cover story of the 16 April 1990 issue of *Time* magazine discussed giant particle accelerators, including the Superconducting Super Collider in Texas, which was ultimately judged to be too expensive for completion. Researchers at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, went on to construct the Large Hadron Collider (LHC) in an existing tunnel. The LHC and other accelerators have been responsible for many major discoveries, but these "colossal colliders" have become colossally costly. Innovative approaches will thus be required to reduce the expense of future colliders in the search for previously unseen particles and physics phenomena. On page 53, the Muon Ionization Cooling Experiment (MICE) collaboration<sup>1</sup> reports results that bring scientists a step closer to realizing one of these innovative approaches: a muon collider.

Muons, like electrons, are elementary particles in the standard model of particle physics, but they have about 200 times the mass of

electrons ([go.nature.com/3twyjb](https://go.nature.com/3twyjb)). This fact has ramifications for the size, and therefore cost, of colliders, and for the energy that can be reached in their particle collisions (and thus their potential for discovery).

Although the goal is to accelerate particles so that they collide at the highest possible energies, the particles actually lose energy through radiation when their trajectories are bent by accelerator magnets. Heavy particles such as protons and muons lose much less energy than do lightweight particles such as electrons. For this reason, the circular colliders that can reach the highest energies (for example, the LHC) use protons. However, protons are not elementary particles. They are made up of elementary particles called quarks, and because the collisions are between bound quarks, only about one-sixth to one-tenth of the energy from proton collisions is available to produce other particles<sup>2</sup>. By contrast, because muons are elementary particles, all of the energy from their collisions is available for particle production.

Muon accelerators would have uses beyond

those for particle colliders. For example, a 'Higgs factory' is a highly desirable facility that would produce huge numbers of elementary particles known as Higgs bosons and allow the properties of these particles to be precisely determined. A Higgs factory based on a conventional linear accelerator that collides electrons and positrons (the antiparticles of electrons) would have to be 10–20 kilometres long<sup>3</sup>. But one based on a circular muon collider would require a circumference of only about 0.3 km (ref. 4). In another example, if muons could be stored in a racetrack configuration that has long, straight sections, the decay of the muons in these sections would produce intense neutrino beams. Such a facility, called a neutrino factory, would shed light on the mysteries of neutrinos and on physics beyond the standard model.

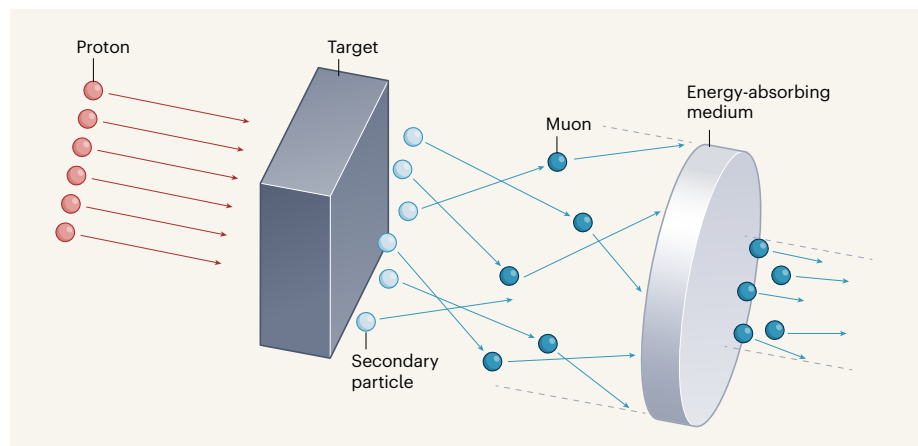
Before a neutrino factory or a muon collider can exist, scientists must learn how to manipulate muon beams. Unlike electron beams, which are produced with almost laser-like quality, muon beams are generated through a complicated process resulting in a beam that is more reminiscent of the spray of pellets from a shotgun. This spray needs to be converted into a laser-like beam.

Such a conversion involves reducing the spread of the muons' positions and velocities in the directions perpendicular to the beam. A temperature can be associated with this spread, and cooling the beam decreases the spread. Several cooling techniques are used at accelerators, but none is fast enough to cool muons, which are unstable and short-lived.

Instead, a method called ionization cooling has been proposed for cooling muon beams<sup>5,6</sup>, although it has never been used. In this approach, muons travel through an accelerator, a portion of which contains a material of low atomic mass, and the spread of the muons' positions and velocities is reduced as the particles ionize atomic electrons in the material. The MICE collaboration's aim was to build and test a system for the ionization cooling of muons, to demonstrate this cooling for the first time and to validate simulation tools for the design of ionization-cooling systems.

In the authors' experiment, a proton beam from the ISIS accelerator at the Rutherford Appleton Laboratory near Didcot, UK, struck a target to produce secondary particles (Fig. 1). Some of these particles decayed into muons, which were directed into an experimental apparatus consisting of focusing magnets, beam instrumentation and a cooling section that contained an energy-absorbing medium made of lithium hydride or liquid hydrogen.

Accelerator experiments usually measure the basic properties of a beam, such as its centre of mass, its spread in particle positions or its density profile. To demonstrate ionization cooling, the MICE collaboration took the unprecedented step of using the technology



**Figure 1 | Production and ionization cooling of muons.** The MICE collaboration<sup>1</sup> carried out an experiment in which a beam of protons was directed at a target to generate secondary particles. Some of these particles decayed into elementary particles known as muons. The positions and velocities of the muons in the resulting beam had a wide spread (indicated by the dashed lines) in the directions perpendicular to the beam. Finally, the muons passed through an energy-absorbing medium made of lithium hydride or liquid hydrogen that reduced this spread by a process called ionization cooling. The process demonstrated by the authors could someday lead to a muon-based particle accelerator.

of collider detectors to measure both the input and output coordinates and velocities of every individual muon that passed through the experimental apparatus. As a result, the authors could unequivocally demonstrate that they had achieved ionization cooling of muons.

Organizations worldwide are developing long-term strategies for exploring the high-energy frontier. Plans include designs for circular colliders up to 100 km in circumference and linear colliders up to 50 km long<sup>7</sup>. Although these approaches, which would use protons or electrons and positrons, have the least technical risk, they still have a substantial cost, as well as technical challenges, that affect their feasibility.

Other plans include designs that would use innovative technologies such as those based on lasers and plasmas<sup>8</sup>. These approaches have made great progress in developing compact accelerator stages at low energy, but the combined use of such stages to reach high energies while retaining a high beam quality will require many years of research and development. Still other plans involve muon beams<sup>9</sup>.

Thanks to the MICE collaboration, the first demonstration of ionization cooling of muons has been achieved. However, it must be noted that the amount of cooling was small. Although conceptual designs for muon colliders have been developed<sup>9</sup>, establishing the viability of a realistic muon-cooling system and of a muon collider will need much more work.

It is too soon to say which, if any, of the proposed approaches will provide a technically and financially feasible path to the future energy frontier. But if physicists can learn how to cool and control muon beams, then it is hard to imagine that putting muons in a circular collider will not be the way forward. These particles offer clean collisions (unlike protons) and lose little energy when their trajectories are bent by accelerator magnets (unlike electrons). As a result, a muon collider could reach energies that match or surpass those of an electron or proton collider, but be substantially smaller. The MICE collaboration's work is a milestone on the road to realistic muon-cooling systems that could someday lead to neutrino factories and muon colliders.

**Robert D. Ryne** is in the Accelerator Technology and Applied Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. e-mail: rdyne@lbl.gov

1. MICE collaboration. *Nature* **578**, 53–59 (2020).
2. Panofsky, W. K. H. & Breidenbach, M. *Rev. Mod. Phys.* **71**, S121–S132 (1999).
3. European Strategy for Particle Physics Preparatory Group. Preprint at <https://arxiv.org/abs/1910.11775> (2019).
4. Rubbia, C. *Int. J. Mod. Phys. A* **33**, 1844010 (2018).
5. Skrinsky, A. N. & Parkhomchuk, V. V. *Sov. J. Part. Nucl.* **12**, 223–247 (1981).
6. Neuffer, D. *Part. Accel.* **14**, 75–90 (1983).
7. Amaldi, U. *et al.* Preprint at <https://arxiv.org/abs/1912.13466> (2019).
8. Katsouleas, T. *Nature* **431**, 515–516 (2004).
9. Boscolo, M., Delahaye, J.-P. & Palmer, M. *Rev. Accel. Sci. Technol.* **10**, 189–214 (2019).

## Climate science

# Early models successfully predicted global warming

Jennifer E. Kay

Climate models published between 1970 and 2007 provided accurate forecasts of subsequently observed global surface warming. This finding shows the value of using global observations to vet climate models as the planet warms.

Climate models are equations that describe climatically relevant processes and are solved on supercomputers. In addition to being invaluable tools for testing scientific hypotheses, these models have long provided societally important forecasts. The first climate models to numerically describe an evolving and interacting atmosphere, ocean and land surface on a grid covering the entire Earth date back to the 1970s (for example, refs 1–3). Since then, the planet's surface has warmed, in large part because of increased emissions of greenhouse gases. Writing in *Geophysical Research Letters*, Hausfather *et al.*<sup>4</sup> retrospectively assessed the forecasting skill of climate models published

between 1970 and 2007. Their results show that the physics in these early models was accurate in predicting subsequently observed global surface warming.

A key point emphasized by the authors is that the forecasting ability of climate models is limited by unknowable future climate drivers. Many major drivers, such as increased concentrations of carbon dioxide in the atmosphere caused by the burning of fossil fuels, result from human activities and decisions. Early climate modellers included estimates for future climate drivers in their forecasts. However, they could not know, for example, how the world would industrialize or the associated



**Figure 1 | A Univac 1108 computer, from 1972.** Hausfather *et al.*<sup>4</sup> demonstrate that climate models published over the past five decades accurately predicted subsequently observed changes in Earth's global mean surface temperature. These models include ones reported in the 1970s that used supercomputers, such as the Univac 1108, that had extremely limited power relative to those used today.



emissions of CO<sub>2</sub> that would result.

Hausfather and colleagues developed a method for evaluating the forecasts of early climate models without penalizing the models for their inaccurate estimates of unknowable future climate drivers. The authors examined 17 projections of global mean surface temperature (GMST) from 14 models. Before applying their method, they found that 10 projections were consistent with observations. But when inaccuracies in the estimates of climate drivers were taken into account, the authors discovered that 14 projections agreed with the data. Of the three that did not, two predicted higher-than-observed surface warming and one predicted lower-than-observed warming.

Developing credible climate models through an understanding of climatically relevant processes, observations and well-formulated equations is a considerable scientific and computational challenge. The equations that describe climate are complex and require substantial computing power to solve. As a result, climate models have always been run on the fastest supercomputers available. It is especially impressive that the earliest models assessed by Hausfather *et al.* produced accurate GMST forecasts, given the extremely limited computing power available then compared with that used today (Fig. 1).

Although the authors' findings show that climate models can accurately predict GMST, these forecasts are insufficient for understanding and preparing for the effects of ongoing climate change. For instance, regional climate change is especially subject to unpredictable climate variability, which greatly limits forecasting potential – even on decadal timescales when the climate drivers are known<sup>5</sup>. Moreover, on the basis of GMST forecasts alone, it is hard to predict, for example: to what extent sea level will rise; how ocean acidification caused by uptake of atmospheric CO<sub>2</sub> will influence marine ecosystems; and the frequency and magnitude of future fires, droughts and floods.

Scientists will have to continue to improve climate modelling and to increase their understanding of the effects of climate change, while keeping in mind the tension between the need for increased model resolution, greater representation of climatically relevant processes, and more simulations to characterize unpredictable climate variability. The successful forecasting of GMST by early climate models is impressive, but leaves much work to be done – as scientists, policymakers and stakeholders are all well aware.

Numerical models based on scientific equations describing the atmosphere are used daily to make decisions that save lives and money. As the climate continues to change owing largely to human activities, scientists need to use, improve and communicate the value of numerical models and the equations and

knowledge that underlie them. Hausfather and colleagues' work demonstrates that the physics in climate models has been providing accurate forecasts of GMST under increasing amounts of atmospheric CO<sub>2</sub> for decades. Such predictions are useful for estimating the maximum amount of CO<sub>2</sub> that can be released into the atmosphere over time to keep surface warming to a specified level.

Crucially, the authors' results also show that a major source of uncertainty in GMST forecasts comes from climate drivers. And, of these drivers, it is emissions of greenhouse gases from human activity that will largely determine future surface warming. The findings indicate the usefulness of climate-model predictions of GMST in response to increasing greenhouse-gas emissions, despite unknowable future climate drivers. But scientists must

also continue to develop climate models in concert with everything else available to them, to plan for a changed climate that requires much more than forecasts of surface warming.

**Jennifer E. Kay** is in the Department of Atmospheric and Oceanic Sciences and at the Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, Colorado 80309, USA.  
e-mail: jennifer.e.kay@colorado.edu

1. Manabe, S. in *Global Effects of Environmental Pollution* (ed. Singer, S. F.) Ch. 3, 25–29 (Springer, 1970).
2. Mitchell, J. M. Jr in *Global Effects of Environmental Pollution* (ed. Singer, S. F.) Ch. 12, 139–155 (Springer, 1970).
3. Benson, G. S. *Proc. Natl Acad. Sci. USA* **67**, 898–899 (1970).
4. Hausfather, Z., Drake, H. F., Abbott, T. & Schmidt, G. A. *Geophys. Res. Lett.* **47**, e2019GL085378 (2020).
5. Deser, C., Phillips, A., Bourdette, V. & Teng, H. *Clim. Dyn.* **38**, 527–546 (2012).

## Cancer

# Brain tumours manipulate neighbouring synapses

Nicola J. Allen

The growth of a brain tumour can be affected by the activity of its neighbouring neurons. The finding that such tumours send signals that boost connections between these neurons reveals a pathway that drives cancer growth. **See p.166**

A type of non-neuronal brain cell called a glial cell can give rise to a lethal cancer called glioblastoma<sup>1</sup>. Half of the cells in the human brain are glial cells, which normally act to support the function and communication of neurons<sup>2</sup>. Yet despite decades of research, there are no existing treatments for glioblastoma that substantially increase the survival time of people with such tumours. On page 166, Yu *et al.*<sup>3</sup> report their analysis of the effects on the brain of certain glioblastoma-associated mutations. These insights might open up new strategies for anticancer research.

DNA sequencing of cancers has identified many tumour-associated mutations. However, it is a challenge to determine which of these mutations have a causal role in tumour development and growth, and which have no effect. Moreover, some mutations can be context dependent, such that the same mutation might have differing effects depending on the type of tumour and its microenvironment. It is hard to predict whether different mutations (variants) of the same gene will lead to the same outcome in different tumour types.

Yu and colleagues tackled these issues by studying the RTK–RAS–PI3K signalling

pathway in glioblastoma. This pathway is altered in 90% of glioblastomas<sup>4,5</sup>, and mutations in it boost cell division and tumour growth. The authors focused on mutations in the gene encoding the enzyme PIK3CA – a pathway component that is often abnormal in human glioblastomas.

The authors generated mouse models of glioblastoma, and mutated genes in the RTK–RAS–PI3K pathway using the gene-editing tool CRISPR–Cas9, which resulted in tumour growth. Yu *et al.* then engineered animals to express PIK3CA variants that are found in human glioblastomas (Fig. 1). This revealed that many of the tested variants made tumours more aggressive and rapidly lethal. This was the case both for known variants and for others that had not previously been associated with a role in glioblastoma.

Yu and colleagues studied whether alterations in the enzymatic activity of PIK3CA variants might explain how they accelerate glioblastoma progression. Surprisingly, the activity of PIK3CA was not always linked to an effect on tumour growth – some variants strongly increased enzyme activity, whereas others had a much milder effect. To investigate other mechanisms that might explain the effect

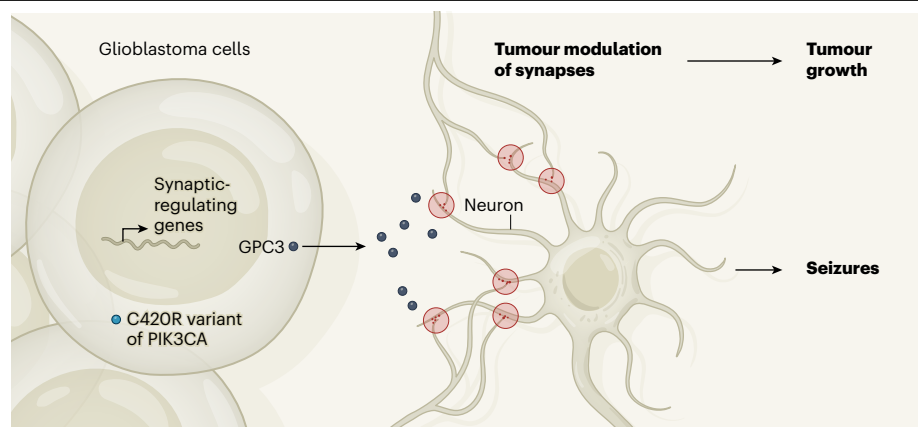
on tumour growth, the authors performed RNA sequencing of tumour cells. Most PIK3CA variants had patterns of altered gene expression that were similar to each other in the animals' tumours, but different from the pattern in healthy tissue. However, two variants (named C420R and H1047R) had distinctive patterns compared with each other and with the other variants. Tumours containing either of these two variants showed abnormally altered expression of hundreds of genes that regulate synapses (structures that connect neurons). C420R and H1047R were also each associated with alterations in different categories of synapse-regulating gene.

Seizures are caused by a rise in neuronal activity and are often an early symptom of glioblastoma<sup>6</sup>. This effect might be driven by an abnormal increase in the connectivity of excitatory synapses (those that drive neuronal activity) and a decrease in the connectivity of inhibitory synapses (which decrease neuronal activity). Yu and colleagues report that mice with tumours that express C420R or H1047R variants had seizures, whereas a PIK3CA variant termed R88Q that does not alter synapse-related gene expression in tumours did not induce seizures. Compared with tumours expressing the R88Q variant, those expressing C420R or H1047R showed an increase in the number of excitatory synapses and a decrease in the number of inhibitory synapses in tissue surrounding the mouse tumours. This pattern might explain how certain PIK3CA variants drive neuronal excitability and seizures.

When the authors engineered mouse brains to express increased amounts of the PIK3CA variants (without expressing other mutations that drive glioblastoma), they found that, in the absence of glioblastoma, H1047R drives seizures whereas C420R and R88Q do not. This indicates that H1047R acts 'cell autonomously' to regulate neuronal excitability – meaning that it functions within the cell itself rather than acting on a neighbouring cell.

To determine how C420R might drive seizures, the authors expressed PIK3CA variants in a type of glial cell called an astrocyte that enhances the formation of synapses between neurons by releasing secreted molecules. The authors cultured astrocytes *in vitro* with neurons and assessed the astrocytes' ability to induce neuronal synapse formation. Astrocytes that expressed C420R induced higher than normal synaptic formation between neurons, whereas astrocytes that expressed H1047R had no such effect. This suggests that C420R changes the properties of glioblastoma cells to make them induce the formation of synapses between the neurons that they contact.

Further analysis of RNA sequencing data by the authors revealed that tumour cells that expressed C420R showed increased expression of known synapse-regulating



**Figure 1 | Tumour modulation of neuronal connections.** Yu *et al.*<sup>3</sup> report their analysis of the brain cancer glioblastoma, which arises from the growth of non-neuronal (glial) cells. Using mouse models, the authors analysed mutations found in people who have glioblastoma. They report that a mutant form of the protein PIK3CA – the C420R variant – is associated with expression of genes in the cancer cell that can regulate the synaptic connections between neurons. One such gene encodes the protein glypican 3 (GPC3), which is secreted by cells and can boost synapse formation (synapses shown in pink). This leads to a rise in synaptic activity, which was associated with tumour progression. Enhanced neuronal activity might explain why seizures are associated with glioblastoma.

factors that are secreted by astrocytes. These included members of the glypican family (part of a group of sugar-containing proteins called proteoglycans), which can induce the formation of excitatory synapses in the brain<sup>7</sup>.

The expression of one member of this family, glypican 3, was particularly highly upregulated in tumour cells that expressed C420R. When the authors engineered glioblastomas expressing C420R to lack glypican 3, this caused a decrease in seizures and a longer lifespan compared with mice that had C420R-expressing glioblastomas that expressed glypican 3. By contrast, animals with glioblastomas engineered to have higher expression of glypican 3 had more seizures and a decreased lifespan compared with animals with glioblastomas that did not overexpress glypican 3. This suggests that glypican 3 is necessary and sufficient in the context of glioblastoma to regulate seizures by controlling synapse formation in the neuronal tissue around the tumour.

Yu and colleagues' work supports growing evidence that tumour cells interact with their neighbouring healthy cells to alter brain function. Recent work has revealed that cells of the glioblastoma itself can form a synapse with surrounding neurons (in such synapses, the tumour has postsynaptic structures and the neuron forms presynaptic structures), and that blocking this synaptic input to the glioblastoma decreases tumour growth<sup>8–10</sup>. Yu *et al.* now reveal how a glioblastoma can remodel connections between its neighbouring neurons, and report that the underlying mechanism differs depending on the specific PIK3CA variant involved. It will be interesting to determine in future studies whether the same synapse-promoting signals are

responsible for regulating the synapses that form between a tumour and its neighbouring neurons, and for regulating synapses that form between neurons surrounding the tumour. This is an intriguing matter, because glypicans regulate neuronal synapse formation through what are known as AMPA receptors, and the synapses that form between neurons and glioblastoma cells use AMPA receptors for signalling<sup>8,9</sup>.

Abnormally high expression of glypican family members occurs in various types of tumour, including liver and pancreatic cancer<sup>11,12</sup>. In those cases, glypicans are thought to bind to and regulate the signalling of growth factors that promote tumour growth. Glypicans are clearly of interest in trying to understand many disorders, and efforts being made to block their function in other cancers might suggest approaches worth testing for use in glioblastoma treatments<sup>11,12</sup>.

**Nicola J. Allen** is at the Salk Institute for Biological Studies, La Jolla, California 92037, USA.  
e-mail: nallen@salk.edu

1. Noorani, I. *Cancers* **11**, 1335 (2019).
2. Allen, N. J. & Lyons, D. A. *Science* **362**, 181–185 (2018).
3. Yu, K. *et al.* *Nature* **578**, 166–171 (2020).
4. Brennan, C. W. *et al.* *Cell* **155**, 462–477 (2013).
5. Cancer Genome Atlas Research Network *Nature* **455**, 1061–1068 (2008).
6. Jung, E. *et al.* *Nature Neurosci.* **22**, 1951–1960 (2019).
7. Allen, N. J. *et al.* *Nature* **486**, 410–414 (2012).
8. Venkataramani, V. *et al.* *Nature* **573**, 532–538 (2019).
9. Venkatesh, H. S. *et al.* *Nature* **573**, 539–545 (2019).
10. Barria, A. *Nature* **573**, 499–501 (2019).
11. Li, N., Gao, W., Zhang, Y.-F. & Ho, M. *Trends Cancer* **4**, 741–754 (2018).
12. Nishida, T. & Kataoka, H. *Cancers* **11**, 1339 (2019).

This article was published online on 29 January 2020.



# An orbital water-ice cycle on comet 67P from colour changes

<https://doi.org/10.1038/s41586-020-1960-2>

Received: 5 August 2019

Accepted: 5 November 2019

Published online: 5 February 2020

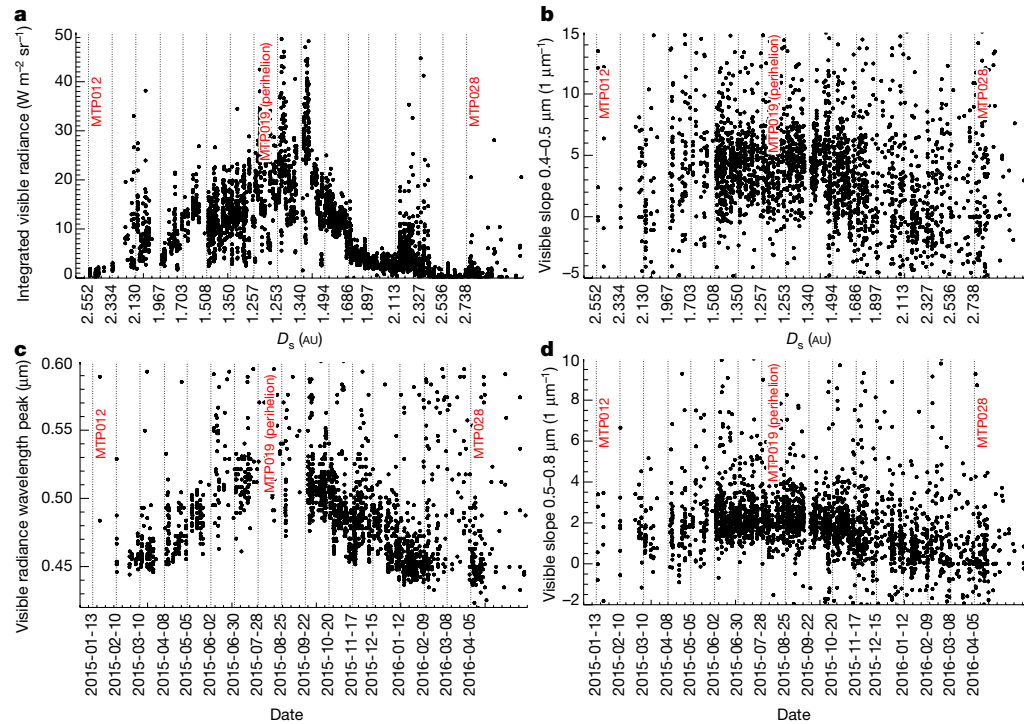
Gianrico Filacchione<sup>1\*</sup>, Fabrizio Capaccioni<sup>1</sup>, Mauro Ciarniello<sup>1</sup>, Andrea Raponi<sup>1</sup>, Giovanna Rinaldi<sup>1</sup>, Maria Cristina De Sanctis<sup>1</sup>, Dominique Bockelée-Morvan<sup>2</sup>, Stéphane Erard<sup>2</sup>, Gabriele Arnold<sup>3</sup>, Vito Mennella<sup>4</sup>, Michelangelo Formisano<sup>1</sup>, Andrea Longobardo<sup>1</sup> & Stefano Mottola<sup>3</sup>

Solar heating of a cometary surface provides the energy necessary to sustain gaseous activity, through which dust is removed<sup>1,2</sup>. In this dynamical environment, both the coma<sup>3,4</sup> and the nucleus<sup>5,6</sup> evolve during the orbit, changing their physical and compositional properties. The environment around an active nucleus is populated by dust grains with complex and variegated shapes<sup>7</sup>, lifted and diffused by gases freed from the sublimation of surface ices<sup>8,9</sup>. The visible colour of dust particles is highly variable: carbonaceous organic material-rich grains<sup>10</sup> appear red while magnesium silicate-rich<sup>11,12</sup> and water-ice-rich<sup>13,14</sup> grains appear blue, with some dependence on grain size distribution, viewing geometry, activity level and comet family type. We know that local colour changes are associated with grain size variations, such as in the bluer jets made of submicrometre grains on comet Hale–Bopp<sup>15</sup> or in the fragmented grains in the coma<sup>16</sup> of C/1999 S4 (LINEAR). Apart from grain size, composition also influences the coma's colour response, because transparent volatiles can introduce a substantial blueing in scattered light, as observed in the dust particles ejected after the collision of the Deep Impact probe with comet 9P/Tempel 1<sup>17</sup>. Here we report observations of two opposite seasonal colour cycles in the coma and on the surface of comet 67P/Churyumov–Gerasimenko through its perihelion passage<sup>18</sup>. Spectral analysis indicates an enrichment of submicrometre grains made of organic material and amorphous carbon in the coma, causing reddening during the passage. At the same time, the progressive removal of dust from the nucleus causes the exposure of more pristine and bluish icy layers on the surface. Far from the Sun, we find that the abundance of water ice on the nucleus is reduced owing to redeposition of dust and dehydration of the surface layer while the coma becomes less red.

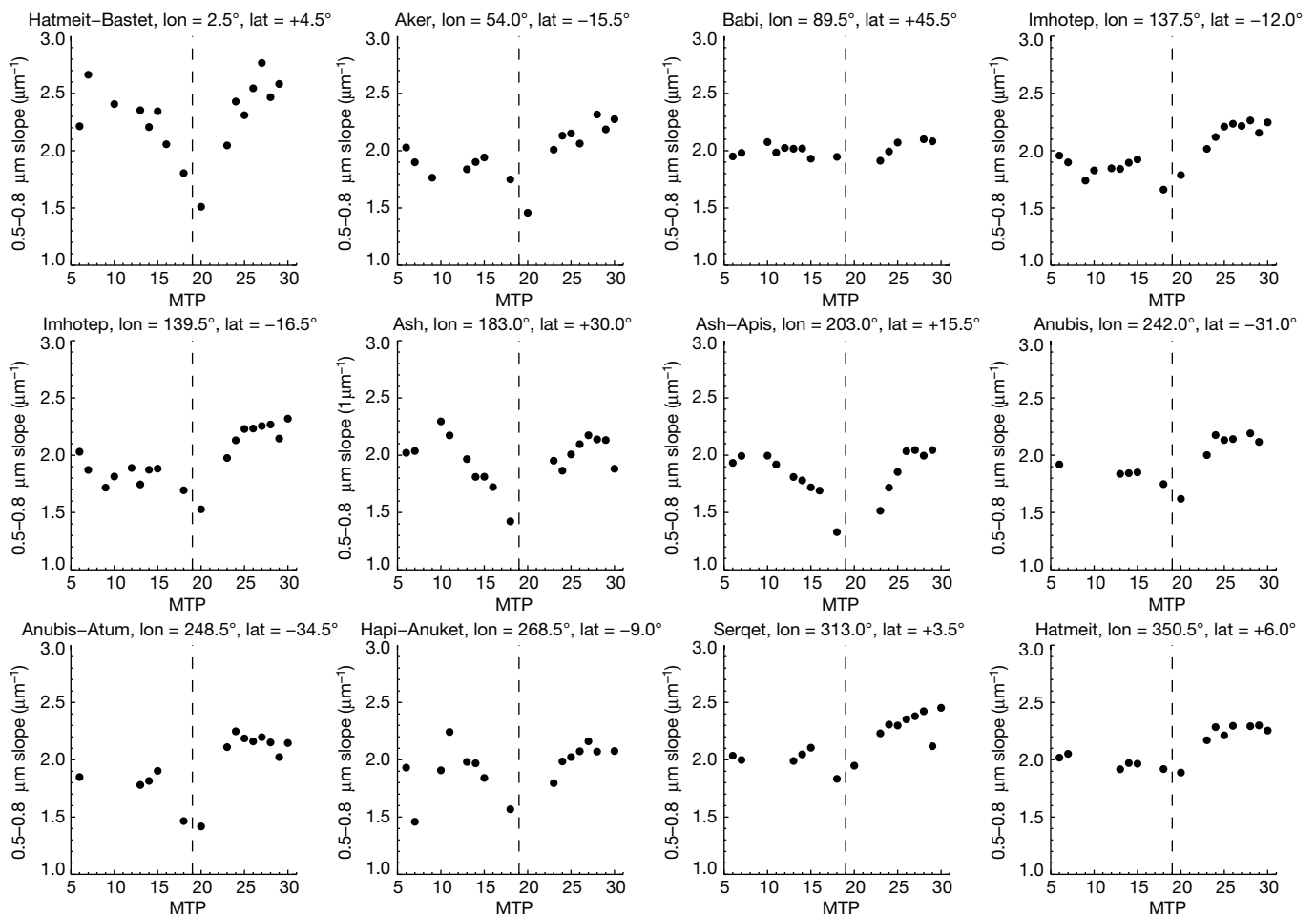
Understanding how comets work and evolve is one of the most compelling questions to which the Rosetta mission<sup>18</sup> has been trying to find an answer. Unlike past fly-by missions to comets, Rosetta was designed to enter the same orbit and accompany comet 67P/Churyumov–Gerasimenko (hereafter 67P) through its perihelion passage, giving us the unique opportunity to follow the colour changes developing on a comet nucleus and coma in its active phase through the inner Solar System. So far, several studies have investigated how dust and gas (H<sub>2</sub>O, CO<sub>2</sub>) production<sup>3</sup> are correlated during one cometary rotation, showing that the water-ice sublimation on the surface of 67P is the driving mechanism for dust ejection in the coma<sup>2</sup>. In fact, the dust emission flux appears lower above high-cohesion consolidated terrains and reaches a maximum when the subsolar direction is aligned above volatile-rich areas. As the illumination conditions continuously change on 67P, rather than focusing on coma and nucleus colour variations occurring during a few diurnal rotations, we analysed the entire dataset

returned by the Visible, Infrared and Thermal Imaging Spectrometer (VIRTIS)<sup>19</sup> on Rosetta. Our method allows exploration of the comet's colour evolution starting from January 2015 (inbound orbit, heliocentric distance 2.55 AU), encompassing perihelion passage (August 2015, 1.24 AU) until May 2016 (outbound orbit, 2.92 AU). Spectral indicators in the visible range, including the integrated radiance (*I*), the wavelength of the radiance peak ( $\lambda_{\text{max}}$ ) and the spectral slopes (in the 0.4–0.5 and 0.5–0.8  $\mu\text{m}$  ranges), are synergistically used to model the composition and grain size distribution of the dust particles in the coma as a function of heliocentric distance. The spectral indicators are derived for each spectrum acquired by VIRTIS in an annulus encircling the nucleus and comprising all the pixels in the coma located at a tangent altitude between 1.0 and 2.5 km above the nucleus' limb. At the same time, we monitor the colour changes occurring on 12 test areas on the nucleus through the 0.5–0.8  $\mu\text{m}$  spectral slope<sup>5,6</sup>. The description of the dataset, including an example of VIRTIS images and spectra after calibration

<sup>1</sup>INAF-IAPS, Institute for Space Astrophysics and Planetology, Rome, Italy. <sup>2</sup>LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université Paris Diderot Sorbonne Paris Cité, Meudon, France. <sup>3</sup>German Aerospace Center (DLR), Institute of Planetary Research, Berlin, Germany. <sup>4</sup>INAF—Osservatorio Astronomico di Capodimonte, Naples, Italy. \*e-mail: gianrico.filacchione@inaf.it



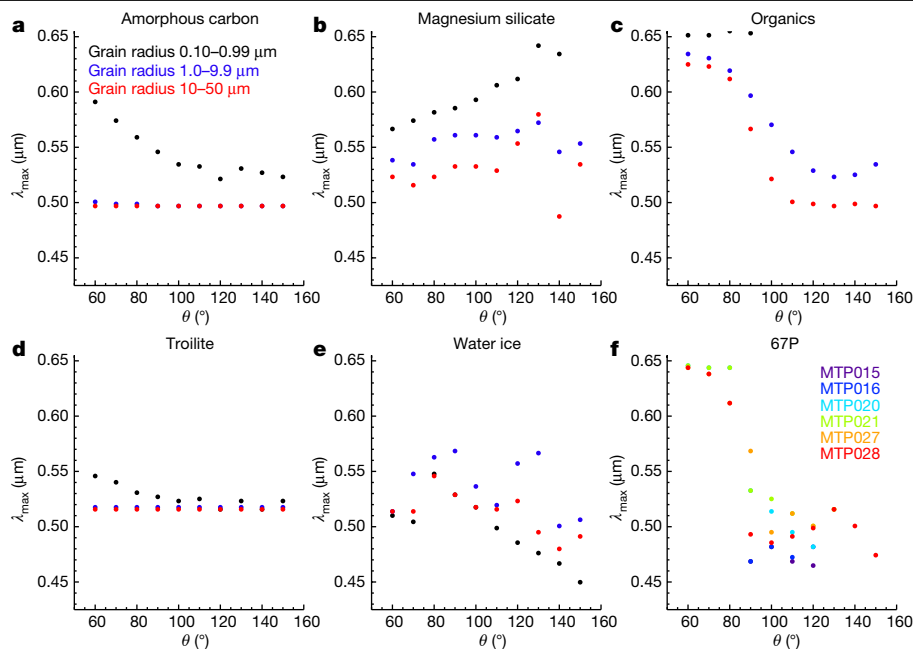
**Fig. 1 | Time series of the spectral properties of 67P coma dust.** **a**, Integrated radiance ( $I$ ); full dataset, partial annulus coverage (black points); reduced dataset, complete annulus coverage (red points; see discussion in Methods). **b**, The 0.4–0.5  $\mu\text{m}$  spectral slope. **c**, Visible radiance wavelength peak  $\lambda_{\text{max}}$ . **d**, The 0.5–0.8  $\mu\text{m}$  spectral slope. Throughout the paper, spectral slopes are given in  $\mu\text{m}^{-1}$  corresponding to 10%/100 nm colour slope. Rosetta's MTP intervals are marked by vertical lines between MTP012 and MTP028. Time intervals and heliocentric distance ( $D_s$ ) of MTP periods are listed in Extended Data Table 1. Perihelion occurs during MTP019. Dates are displayed as year-month-day.



**Fig. 2 | Time evolution of 67P nucleus colour measured through the 0.5–0.8  $\mu\text{m}$  spectral slope above 12 control areas.** Perihelion passage occurs during MTP019 and is marked by the vertical dashed line. The morphological

region name and position (longitude, lon, and latitude, lat) of each area are reported above each plot. The maximum error associated with the slope measurements is of the order of  $\pm 0.1 \mu\text{m}^{-1}$ .





**Fig. 3 | Simulations of the  $\lambda_{\max}$  as a function of the scattering angle for spherical particles of different composition. a, Amorphous carbon. b, Magnesium silicate. c, Organic ice tholin ( $\lambda_{\max} > 0.65 \mu\text{m}$  for 0.10–0.99- $\mu\text{m}$ -radius grains). d, Troilite. e, Water ice. f, 67P observations by VIRTIS. Median**

data computed on  $10^\circ$  bin scattering angles ( $\theta$ ) during different mission phases. Standard deviation computed on VIRTIS data are of the order of  $0.03 \mu\text{m}$ .

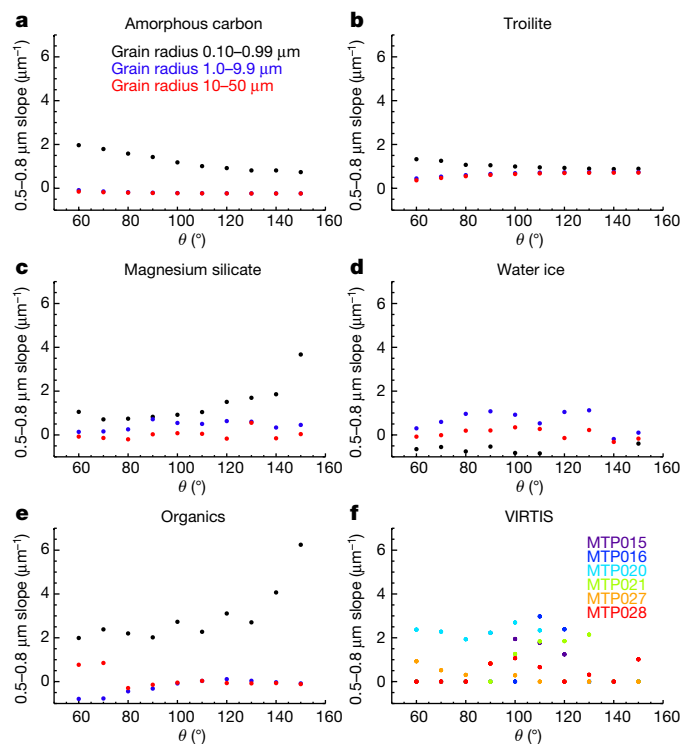
improvement, spectral indicator processing and data modelling, are given in Methods.

From a synergistic time analysis of the coma and of the nucleus surface spectral properties, two opposite trends with heliocentric distance are observed. Coma integrated radiance increases (see  $I$  time series in Fig. 1a) and colour becomes redder when the comet approaches perihelion (see time series of  $\lambda_{\max}$  and the 0.4–0.5 and 0.5–0.8  $\mu\text{m}$  spectral slopes in Fig. 1b–d). Conversely, on 12 control areas of the nucleus, for which we have continuous time coverage during the entire Rosetta mission, we observe a systematic blueing of the surface (see 0.5–0.8  $\mu\text{m}$  spectral slope time series in Fig. 2), with a reduction of the spectral slope up to 50% measured between the passage of 67P through the frost line (heliocentric distance 2.7 AU, occurring in October 2014 during inbound orbit and in April 2016 on the outbound) and perihelion (1.25 AU, August 2015). During the pre-perihelion period, we interpret the coma colour change as being the consequence of the progressive loss of the ice fraction in the dust grains ejected from the nucleus<sup>20</sup>, which makes them redder.

When the comet approaches the Sun, the dust grains lifted from the nucleus' southern hemisphere—the one illuminated at that time—are warmer and become more dehydrated. Simulations based on Mie scattering theory applied to spherical grains of radius between 0.1 and  $50 \mu\text{m}$  and five different compositions (see Methods for a discussion about the scattering model limitations induced by spherical shaped grains) are used to model VIRTIS spectral indicators, for example,  $\lambda_{\max}$  and the 0.5–0.8  $\mu\text{m}$  spectral slope. Simulation results (Figs. 3, 4) are compatible with the presence of submicrometre grains made of organic matter and carbon during perihelion passage, which could be the result of the fragmentation of the mineral and organic matrix embedded in larger grains. During the outbound orbit, the simulations show that the organic material continues to be present in the coma together with an increasing percentage of different grains (water ice or magnesium silicate, both of which are not fully compatible with the spectral indicator trends (Extended Data Table 3)), contributing to the progressive blueing observed in the data. In this orbital phase, the Sun shines again on the northern hemisphere, from where a greater flux of surficial water-ice-rich grains<sup>21</sup> is lifted in the coma.

Conversely, the nucleus' surface becomes progressively bluer during the inbound orbit because the greater solar input causes an increase in the gaseous activity. This triggers dust ejection with consequent erosion of the surface: on average, a 0.5-m-thick erosion layer is lost during each orbit<sup>21</sup>. As a result of this process, more pristine subsurface layers enriched in water ice are exposed on the surface, which turns bluer<sup>5,6,22</sup>. We have evidence<sup>6</sup>, in fact, that water-ice-rich layers are immediately available under the dust layer, as shown in areas of recent disruption of the surface<sup>23</sup> and as demonstrated by many thermal models applied to cometary nuclei<sup>24,25</sup>. Between 20% (ref. <sup>21</sup>) and 50% (ref. <sup>26</sup>) of the dust grains ejected around perihelion from the south pole—the region that receives the maximum insolation at that time—fall back preferentially on the northern hemisphere. The fall-back flux is dominated by decimetre-size dust aggregates in which a small fraction of water ice can be maintained<sup>21</sup>. After perihelion, as soon as activity begins to settle, the progressive accumulation of dust on the surface covers the exposed pristine layers again, making the nucleus surface redder again. This process explains the colour cycle observed by VIRTIS on the nucleus (Fig. 2).

The colour of the 67P coma particles can be modelled with different grain populations made of water ice before perihelion, organic material and carbon at perihelion, and water ice or possibly magnesium silicate after perihelion. Similar composition endmembers are compatible with the findings reported by previous Rosetta studies: during outbursts, the release of submicrometre-size water-ice grains mixed with hundred-micrometre-size refractory grains has been observed<sup>27</sup>. At the same time, a broad emission feature between 3.4 and  $3.6 \mu\text{m}$  has been detected by VIRTIS<sup>28</sup>, which is compatible with the sublimation of organic materials caused by the high temperatures of the grains. The colour reduction observed when the comet was far from the Sun resembles the spectral properties of water-ice grains, whose existence in the lower coma has been confirmed by other investigations: apart from the presence of submicrometre-size water-ice grains ejected during outbursts<sup>27</sup>, thermal models and Optical, Spectroscopic, and Infrared Remote Imaging System (OSIRIS) camera images<sup>20</sup> show that at 1.53 AU heliocentric distance, the total sublimation of ice for grains of



**Fig. 4 | Simulations of the 0.5–0.8  $\mu\text{m}$  spectral slope as a function of the scattering angle for spherical particles of different composition. **a**, Amorphous carbon. **b**, Troilite. **c**, Magnesium silicate. **d**, Water ice. **e**, Organics. **f**, Median spectral slope derived from VIRTIS observations for pre-perihelion (MTP015 and MTP016), perihelion (MTP020 and MTP021) and post-perihelion (MTP027 and MTP028) mission phases. The standard deviations of the VIRTIS data are  $\leq 0.6 \mu\text{m}^{-1}$ .**

radius between 5 and 500  $\mu\text{m}$  made of ice-dust layers or mixtures occurs well beyond the 2.5 km distance from the 67P nucleus considered here. This implies that water-ice grains can populate the region of the coma analysed in this study. Finally, a photometric investigation of the visible colours of more than 1,000 individual grains resolved on OSIRIS colour images of the 67P coma<sup>29</sup> indicates the presence of three composition classes: organic material grains (steep colour slope), mixtures of silicate and organic material (intermediate slope), and water ice (flat slope). The presence of carbon, silicates and magnesium in 67P dust grains has been assessed by the Cometary Secondary Ion Mass Analyser (COSIMA) instrument on Rosetta<sup>30</sup>.

The fact that the coma's and the surface's colour spectral indicators return to the same values at the beginning and at the end of their time series (Fig. 1), for example, at heliocentric distances greater than 2.7 AU, is evidence of the establishment of an orbital water-ice cycle driven by the solar heating. Far from the Sun, with the settling down of the gaseous activity<sup>3</sup>, a tenuous ( $I \approx 0$  in Fig. 1a) coma surrounds a red-coloured dusty nucleus.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1960-2>.

- Gundlach, B., Blum, J., Keller, H. U. & Skorov, Y. V. What drives the dust activity of comet 67P/Churyumov-Gerasimenko? *Astron. Astrophys.* **583**, A12 (2015).
- Tubiana, C. et al. Diurnal variation of dust and gas production in comet 67P/Churyumov-Gerasimenko at the inbound equinox as seen by OSIRIS and VIRTIS-M on board Rosetta. *Astron. Astrophys.* **630**, A23 (2019).
- Hansen, K. C. et al. Evolution of water production of 67P/Churyumov-Gerasimenko: an empirical model and a multi-instrument study. *Mon. Not. R. Astron. Soc.* **462**, S491–S506 (2016).
- Bockelée-Morvan, D. et al. VIRTIS-H observations of comet 67P's dust coma: spectral properties and color temperature variability with phase and elevation. *Astron. Astrophys.* **630**, A22 (2019).
- Filacchione, G. et al. The global surface composition of 67P/CG nucleus by Rosetta/VIRTIS. (I) Prelanding mission phase. *Icarus* **274**, 334–349 (2016).
- Ciarniello, M. et al. The global surface composition of 67P/Churyumov-Gerasimenko nucleus by Rosetta/VIRTIS. (II) Diurnal and seasonal variability. *Mon. Not. R. Astron. Soc.* **462**, S443–S458 (2016).
- Güttler, C. et al. Synthesis of the morphological description of cometary dust at comet 67P. *Astron. Astrophys.* **630**, A24 (2019).
- Huebner, W. F. et al. (eds) *Heat and Gas Diffusion in Comet Nuclei*, SR-004, June, 2006 (ESA Publications Division, 2006).
- Lauter, M. et al. Surface localization of gas sources on comet 67P/Churyumov-Gerasimenko based on DFMS/COPS data. *Mon. Not. R. Astron. Soc.* **483**, 852–861 (2019).
- Jewitt, D. & Meech, K. J. Cometary grain scattering versus wavelength, or, "What color is comet dust?" *Astron. Astrophys. J.* **310**, 937–952 (1986).
- Zubko, E. et al. Interpretation of photopolarimetric observations of comet 17P/Holmes. *J. Quant. Spectrosc. Radiat. Transf.* **112**, 1848–1863 (2011).
- Hadamcik, E. et al. Linear polarization of light scattered by cometary analogs: new samples in *Asteroids, Comets and Meteors 2014* (eds Muinonen, K. et al.) (2014).
- Beer, E. in *Deep Impact as a World Observatory Event: Synergies in Space, Time and Wavelengths* (eds Käufel, H. U. & Sterken, C.) 59–67 (Springer, 2009).
- Fernández, Y. R. et al. Near-infrared light curve of comet 9P/Tempel 1 during Deep Impact. *Icarus* **187**, 220–227 (2007).
- Furusho, R. et al. Imaging polarimetry and color of the inner coma of comet Hale-Bopp (C/1995 O1). *Publ. Astron. Soc. Jpn* **51**, 367–373 (1999).
- Hadamcik, E. & Levasseur-Regourd, A. C. Dust coma of comet C/1999 S4 (LINEAR): imaging polarimetry during nucleus disruption. *Icarus* **166**, 188–194 (2003).
- Hodapp, K. W. et al. Visible and near-infrared spectrophotometry of the Deep Impact ejecta of comet 9P/Tempel 1. *Icarus* **187**, 185–198 (2007).
- Taylor, M. G. G. T., Altobelli, N., Buratti, B. J. & Choukroun, M. The Rosetta mission orbiter science overview: the comet phase. *Phil. Trans. R. Soc. A* **375**, 20160262 (2017).
- Coradini, A. et al. VIRTIS: an imaging spectrometer for the Rosetta mission. *Space Sci. Rev.* **128**, 529–559 (2007).
- Gicquel, A. et al. Sublimation of icy aggregates in the coma of comet 67P/Churyumov-Gerasimenko detected with the OSIRIS cameras on board Rosetta. *Mon. Not. R. Astron. Soc.* **462**, S57–S66 (2016).
- Keller, H. U. et al. Seasonal mass transfer on the nucleus of comet 67P/Churyumov-Gerasimenko. *Mon. Not. R. Astron. Soc.* **469**, S357–S371 (2017).
- Fornasier, S. Rosetta's comet 67P/Churyumov-Gerasimenko sheds its dusty mantle to reveal its icy nature. *Science* **354**, 1566–1570 (2016).
- Filacchione, G. et al. Exposed water ice on the nucleus of comet 67P/Churyumov-Gerasimenko. *Nature* **529**, 368–372 (2016).
- De Sanctis, M. C., Capria, M. T. & Coradini, A. Thermal evolution model of 67P/Churyumov-Gerasimenko, the new Rosetta target. *Astron. Astrophys.* **444**, 605–614 (2005).
- Capria, M. T. et al. How pristine is the interior of the comet 67P/Churyumov-Gerasimenko? *Mon. Not. R. Astron. Soc.* **469**, S685–S694 (2017).
- Hu, X. et al. Seasonal erosion and restoration of the dust cover on comet 67P/Churyumov-Gerasimenko as observed by OSIRIS onboard Rosetta. *Astron. Astrophys.* **604**, A114 (2017).
- Agarwal, J. et al. Evidence of subsurface energy storage in comet 67P from the outburst of 2016 July 03. *Mon. Not. R. Astron. Soc.* **469**, s606–s625 (2017).
- Bockelée-Morvan, D. et al. Comet 67P outbursts and quiescent coma at 1.3 au from the Sun: dust properties from Rosetta/VIRTIS-H observations. *Mon. Not. R. Astron. Soc.* **469**, S443–S458 (2017).
- Frattini, E. et al. Post-perihelion photometry of dust grains in the coma of 67P Churyumov-Gerasimenko. *Mon. Not. R. Astron. Soc.* **469**, S195–S203 (2017).
- Bardyn, A. et al. Carbon-rich dust in comet 67P/Churyumov-Gerasimenko measured by COSIMA/Rosetta. *Mon. Not. R. Astron. Soc.* **469**, S712–S722 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

## Methods

### VIRTIS dataset

While the scientific objectives of the Rosetta mission<sup>13,18</sup>, including the study of the dust in the coma, were well-defined before the spacecraft reached 67P, the detailed planning of the scientific payload operations has been a complex and continuously evolving process due to the uncertainties of the cometary environment in which the mission had to operate. A detailed overview of the Rosetta science planning process and challenges is given in ref. <sup>31</sup>. In this context, VIRTIS coma observations were performed from very different viewing and illumination geometries, spacecraft ranges from the nucleus, heliocentric distances and cometary activity levels, resulting in a dataset showing a very variable spatial resolution and signal-to-noise ratio.

While VIRTIS-led targeted observations of the coma have been routinely executed, a large fraction of the VIRTIS dataset consisted of ride-along and serendipitous acquisitions that were acquired while the spacecraft attitude and pointing were controlled by other instruments. For this reason, to maximize the retrieval of information about the dust particles properties, a statistical approach has been applied to the vast VIRTIS dataset with the scope to exploit it in its completeness.

We restrict our analysis to only the VIRTIS-M (imaging channel) visible data covering the 0.25–1.0  $\mu\text{m}$  spectral range with 432 bands and a spectral sampling of less than 2 nm per band. A detailed description of the VIRTIS experiment is provided by ref. <sup>19</sup>. In contrast to the VIRTIS-M infrared channel operating within the 1–5  $\mu\text{m}$  spectral range, which ceased operation in May 2015 due to a permanent failure of the cryo-cooler cooling of the infrared channel detector, the visible channel has operated flawlessly for the entire duration of the mission.

For this study, we processed coma data collected during Rosetta's MTP012–MTP028 periods. The time intervals of each medium-term phase (MTP) are listed in Extended Data Table 1. The dataset consists of more than 4,500 hyperspectral cubes of the coma of 67P acquired from a range of distances from the nucleus between 27.8 and 1,462.2 km, resulting in a spatial resolution between 7 and 365 m per pixel, respectively.

Each VIRTIS observation has been calibrated in spectral radiance<sup>32</sup> and specific geometry parameters have been calculated for each pixel falling on the coma region and the nucleus. Further details about the VIRTIS data mapping methods are given in ref. <sup>5</sup>. The nucleus solar phase and the spacecraft position and distance in the Cartesian frame centred on the nucleus have been calculated using SPICE kernels<sup>33</sup>.

The spectral parameters used in our analysis to study the coma emission are calculated in an annulus defined by a tangent altitude running between  $a_{\min} = 1$  km and  $a_{\max} = 2.5$  km around the nucleus. This criterion is adopted independently from the nucleus–spacecraft distance and the solar phase angle at the time of the observation. The selection of pixels placed at distances larger than 1 km allows (1) reduction of the contribution of the instrumental stray light coming from the illuminated part of the nucleus and (2) maximization of the coverage across the dataset.

An example of a VIRTIS-M visible image of the coma showing the annulus definition is given in Extended Data Fig. 1 (top panels). This particular image refers to observation MTP019–STP068–V1\_00397724286 acquired on 9 August 2015 from a distance of 304 km, resulting in a spatial resolution of 75 m per pixel with an integration time of 16 s per line.

Such a viewing geometry is not very common because the completeness of the annulus over the 360° azimuth (up to distance  $a_{\max}$ ) is limited by the instrumental field of view (FOV = 3.6°), by the relative distance ( $d$ ) between the Rosetta spacecraft and the comet's surface, and by the spacecraft off-nadir pointing direction ( $\alpha$ ). Only when the following condition is verified, VIRTIS-M can acquire the full annulus:

$$\alpha + \arctan\left(\frac{a_{\max} + r}{d}\right) \leq \text{FOV} \quad (1)$$

where  $r = 1.74$  km is the average radius of the nucleus<sup>34</sup>. A similar condition is verified in only 248 observations (marked as red points in Fig. 1a),

corresponding to about 5% of the dataset: as a consequence, integrated radiance and maximum emission wavelength values could be biased by the incompleteness of the annulus on the remaining observations. The average integrated radiance on incomplete annulus observations (black points) appears, in general, higher than on the ones where the annulus is entirely acquired (red points). This happens because VIRTIS is preferentially observing the coma towards the solar direction to maximize the signal-to-noise ratio. Apart from the above limitations, which introduce some scattering in the data points, the value of the integrated radiance is mainly driven by the number of scatterers (dust grains) along the line of sight, rather than by their composition and grain size distribution (which the spectral slopes and the spectral position of the radiance peak depend on).

The overall coma brightness and spectral properties measured by VIRTIS are driven by light scattering on dust particles and are not altered by emissions from gas molecules in the coma. The instrument sensitivity at visible wavelengths, in fact, is not enough to measure gaseous emissions from CN, C<sub>2</sub> and C<sub>3</sub>.

### Calibration update using stellar observations

With the aim of improving the VIRTIS-M visible response in the 0.25–0.45  $\mu\text{m}$  spectral range, where the standard pipeline suffers large uncertainties due to stray light introduced by the on-ground calibration setup, we applied a radiometric correction based on stellar observations. The average signal derived from Vega ( $\alpha$  Lyrae) observations listed in Extended Data Table 2 is used as a reference to correct the standard responsivity. Four observations of Vega acquired with the longest possible integration time, 50 s, were used to retrieve the raw signal above the noise level (about  $\pm 2$  data numbers (DN)) estimated on the nearby deep-sky pixels.

Before this step, each observation was dark-level-subtracted and despiked (Extended Data Fig. 2 (top-left panel), showing the signal and sky level for one of the four Vega observations). Owing to the instrumental point spread function and spectral tilt, both described in detail in ref. <sup>35</sup>, the Vega spectral signal is distributed across 18 spatial pixels and is averaged above these. The four observations were processed in the same way before being further averaged to improve the signal-to-noise ratio. The resulting average Vega signal is shown in Extended Data Fig. 2 (top-right panel). The fractional deviation of the single spectra from the average is about 0.05. The Vega flux in ref. <sup>36</sup> is convolved with the VIRTIS Gaussian response (full-width at half-maximum, 2.3 nm) and then by a 4-nm-wide boxcar. The spectra were then placed on an absolute wavelength scale using the centre of Balmer and Paschen intense absorption lines as references and assuming a constant dispersion. The spectral dispersion (in nm) applied is  $\lambda(b) = 232.9 + 1.88515 \times b$ , with  $b$  ranging from 0 to 431, corresponding to the VIRTIS-M number of spectral bands. The resulting relative responsivity derived as the ratio between VIRTIS signal and Vega flux is shown in Extended Data Fig. 2 (bottom-left panel). To further remove residual high-frequency noise due to the low Vega signal, a nine-band running boxcar filter was applied. The response curves were normalized at the value of 58.75 at  $\lambda = 0.635 \mu\text{m}$  to allow a direct comparison with the standard pipeline responsivity shown in Extended Data Fig. 2 (bottom-right panel). Within the limits of the low Vega signal in the blue spectral range (less than 40 DN at 0.55  $\mu\text{m}$ ), this method allows the retrieval of a better responsivity at shorter wavelengths with respect to the standard pipeline<sup>35</sup>. As shown by the ratio plot (blue curve), the standard pipeline overestimates the target radiance for wavelengths  $\lambda < 0.4 \mu\text{m}$ . In the remaining spectral range, the differences are much smaller. As the wavelength of the dust emission spectral radiance peak occurs between 0.45 and 0.55  $\mu\text{m}$ , such an instrument-response correction allows improvement of the retrieval of the left wing of the radiance peak and the spectral slope determinations.

### Coma spectral indicators

After averaging the coma signal within the annulus previously defined and applying the updated responsivity function, we calculated four



spectral indicators for each observation. The indicators are as follows. (1) The integrated radiance ( $I$ ) across the visible spectral range, defined as:

$$I = \frac{\sum_{n=1}^N \int_{0.25\mu\text{m}}^{1\mu\text{m}} R(n, \lambda) d\lambda}{N} \quad (1)$$

where  $R(n, \lambda)$  is the spectral radiance measured on the  $n$ th pixel of the annulus at wavelength  $\lambda$  and  $N$  is the total number of pixels within the annulus area. (2) The wavelength of maximum emission of the radiance ( $\lambda_{\text{max}}$ ) measured on a fourth-degree fit on the average spectral radiance calculated on the annulus.

After converting spectral radiance to irradiance/solar flux ( $I/F$ ), we calculated the two spectral slopes following ref. <sup>5</sup>. These are: (3) 0.4–0.5  $\mu\text{m}$  spectral slope on  $I/F$ ; and (4) 0.5–0.8  $\mu\text{m}$  spectral slope on  $I/F$ . The two spectral slopes are calculated as the angular coefficient of the best linear fit in the ranges 0.4–0.5 and 0.5–0.8  $\mu\text{m}$  after having normalized  $I/F$  at 0.5  $\mu\text{m}$ . The determination of the spectral indicators follows the scheme shown in Extended Data Fig. 1 (bottom panels).

Spectral indicators depend on illumination geometries and the spacecraft–nucleus distance: as a consequence of the demanding mission scenario, complex orbits, including terminator, icosahedron, flybys and far excursions, have been flown<sup>31</sup>, resulting in a wide range of distances and solar phases. The correlation between the integrated radiance  $I$  and distance is shown in Extended Data Fig. 3, where VIRTIS-M measurements are co-located with Rosetta’s position and solar phase angle with respect to the comet nucleus along the flight trajectory.

## Temporal evolution of the coma colour

The integrated radiance ( $I$ ) time series (Fig. 1a) shows an increase in the coma’s brightness starting from January 2015 (MTP012, heliocentric distance 2.55 AU) to perihelion passage (MTP019, 1.24 AU, August 2015). Owing to the characteristics of Rosetta’s orbit around the nucleus and the different observation strategies implemented during the mission, the completeness of the annulus region on the VIRTIS dataset is reached on only a limited number of observations, marked as red points in Fig. 1a. For the majority of the remaining observations (black points), partial coverage of the annulus is achieved. Notwithstanding the relative scatters of the data introduced by the variability of the observing conditions, some clear trends are identified. A local maximum,  $25 \leq I \leq 30 \text{ W m}^{-2} \text{ sr}^{-1}$ , is measured about one month after perihelion. Two additional intensity peaks are measured at the end of August and September 2015, when a maximum intensity of about  $50 \text{ W m}^{-2} \text{ sr}^{-1}$  is reached. These peaks are caused by energetic cometary activity—two outbursts observed by VIRTIS on 13–14 September 2015 are reported by refs. <sup>28,37</sup>—but appear also to be strongly correlated with the peculiar viewing geometry and spacecraft position with respect to the nucleus: in fact, at the time of these observations, Rosetta was performing far-nucleus excursions, orbiting at greater distances than usual from the nucleus, up to 450 and 1,462 km, respectively. As a consequence of the increased distance from the nucleus, VIRTIS-M has recorded higher radiances boosted by the larger column density of dust particles along the line of sight and by the maximum activity occurring immediately after perihelion. After this period, the integrated radiance drops rapidly from MTP021 to MTP024 where an average value of  $5 \text{ W m}^{-2} \text{ sr}^{-1}$  is reached. The maximum integrated radiance measured in MTP020 and MTP021 is about a factor of 10 more intense than the average value of  $5 \text{ W m}^{-2} \text{ sr}^{-1}$  measured when the comet was far from the Sun (greater than 2 AU) on both inbound and outbound legs of the orbit. With the exclusion of the two peaks, a similar asymmetric, or cusp-like trend with a maximum occurring about one month after perihelion, has been measured by other Rosetta’s instruments for the water production rate, which is associated with the ejection of the dust<sup>3</sup>. Apart from the greater column density, we have clues that a larger particle albedo could contribute to the increase of the integrated radiance: according

to ref. <sup>4</sup>, the albedo of the dust particles increases up to 20% near perihelion for observations acquired at a 90° solar phase angle. Similar albedo changes could be the result of a different composition of the dominant grain population in the coma along the orbit, as discussed in the main text.

The time series of the wavelength of maximum emission ( $\lambda_{\text{max}}$ ) (Fig. 1c) shows a cusp-like trend starting from low values dispersed around  $\lambda_{\text{max}} = 0.45 \mu\text{m}$  on January 2015 (MTP012) on the inbound leg of 67P trajectory to  $\lambda_{\text{max}} > 0.5 \mu\text{m}$  during perihelion passage (MTP019). Moving along the outbound leg of the orbit, the wavelength of maximum emission shifts progressively towards shorter wavelengths, reaching again  $\lambda_{\text{max}} = 0.45 \mu\text{m}$  in mid-May 2016 (MTP028). Whereas  $\lambda_{\text{max}}$  and integrated radiance time series show a similar time trend, spectral slopes time series (Fig. 1b, d) are characterized by a different evolution: at the two extremes of the time series, both 0.4–0.5  $\mu\text{m}$  and 0.5–0.8  $\mu\text{m}$  slopes are almost neutral whereas at perihelion they reach values dispersed at 5 and 3  $\mu\text{m}^{-1}$ , respectively. The visible colours of the coma show a systematic reddening before perihelion passage and a progressive blueing after it. The concurrent variations occurring on all spectral indicator time series point to a change in the dust grain composition and grain size distribution with the heliocentric distance.

## Temporal evolution of the nucleus colour

When observed at a global scale, 67P surface dust appears very dark (geometric visible albedo 6.2%; ref. <sup>38</sup>) and dehydrated, with an average water-ice fraction of about 1% dispersed in submicrometre grains<sup>39,40</sup>. The residual water-ice content, if present, on the dust grains ejected from the surface undergoes rapid sublimation once the grain is lifted<sup>20</sup>. So far, infrared spectroscopic observations have confirmed the presence of only two ices on the surface of 67P nucleus: water ice<sup>23,41</sup> and carbon dioxide<sup>42</sup>.

The low thermal capacity of the dust ( $700 \text{ J kg}^{-1} \text{ K}^{-1}$ ) and the high solar flux can rapidly increase the temperature of the grains causing the sublimation of the volatiles. A colour temperature as high as 630 K is measured on dust grains during outbursts<sup>28</sup> and in the 260–320 K range on quiescent coma<sup>4</sup>. Furthermore, the grains observed by Rosetta’s dust instruments appear dehydrated<sup>30,43</sup>. This probably happens because the instruments aboard the spacecraft are not kept at cryogenic temperatures, allowing the dispersion of the volatile fraction of the grains after their collection. Besides dehydration, grain size distribution also plays a role in the changes of the visible colours (see ‘Modelling dust grain composition and size distribution in the coma’).

During the pre-perihelion phase, the nucleus’ surface shows a progressive blueing until one month after perihelion when the minimum slope is reached in MTP020 (Fig. 2). In this phase, the gas activity removes the dust from the surface with greater efficiency, resulting in the exposure of more pristine subsurface material enriched in ices<sup>5,6,22</sup>. This trend is interrupted after the perihelion passage when a progressive surface reddening is observed. Along the outbound trajectory, the gaseous activity progressively settles, resulting in the accumulation of dehydrated (red) dust on the surface. This appears to be a general mechanism, occurring on the majority of the nucleus surface, as shown in Fig. 2, where the average 0.5–0.8  $\mu\text{m}$  slope is shown for twelve  $30 \text{ m} \times 30 \text{ m}$ -wide control areas at one-month time resolution. These control areas, selected between latitude +45.5° and –34.5°, are the ones that have been observed in daylight by VIRTIS during the entire duration of the mission, allowing us to follow the seasonal colour cycle developing on the nucleus.

## Dust-scattering simulations

The simulations are computed following the Bohren–Huffman Mie scattering code<sup>44</sup> to calculate scattering and absorption by a homogenous isotropic sphere. The method relies on the following assumptions. (1) Single-scattering approximation, for example, the solar photons interact with only one particle before reaching the observer. Such a

hypothesis is reasonable for the annulus region considered in this study, where the optical depth is small. (2) Polydisperse spherical grain size distributions with radii of 0.10–0.99  $\mu\text{m}$  (90 bins, 0.01  $\mu\text{m}$  per bin), 1.0–9.9  $\mu\text{m}$  (90 bins, 0.1  $\mu\text{m}$  per bin) and 10–50  $\mu\text{m}$  (40 bins, 1  $\mu\text{m}$  per bin). (3) Five different grain compositions are simulated: water ice (optical constants from ref. <sup>45</sup>), organic material<sup>46</sup>, troilite<sup>45</sup>, amorphous carbon<sup>47</sup> and magnesium silicate<sup>48</sup>. These endmembers are representative of the cometary nuclei composition, which is made up of a macromolecular assemblage in which various organic components (both aromatic and aliphatic), minerals (including silicates, iron sulphides like pyrrhotite and/or troilite, and possibly ammoniated salt) and ices are mixed together<sup>49</sup>. (4) Spectral simulations are performed at 27 wavelengths within the 0.35–1.0  $\mu\text{m}$  spectral range. (5) Scattering angle  $\theta$  ( $= 180^\circ - g$ , where  $g$  is the solar phase angle) is computed at a step of  $1^\circ$  from  $55^\circ$  to  $155^\circ$  and then averaged on  $4^\circ$  bins to match the VIRTIS FOV ( $3.66^\circ$ ).

Following ref. <sup>50</sup>, the single scattering intensity is given by:

$$I(\lambda) = S(\lambda)n_{\text{col}}(\rho)\sigma q(\lambda) \frac{p(g, \lambda)}{4\pi} \quad (2)$$

where  $S(\lambda)$  is the solar flux<sup>51</sup> scaled at the comet's heliocentric distance,  $n_{\text{col}}$  is the dust column density calculated along the line of sight from an observer placed at distance  $\rho$  from the nucleus,  $\sigma$  is the grain geometrical cross-section,  $q(\lambda)$  is the scattering efficiency of a single particle and  $p(g, \lambda)$  is the scattering phase function calculated at phase  $g$ . Rather than modelling the absolute value of the intensity, we limit here our analysis to the changes occurring in the spectral response of the average radiance emitted from the dust. In particular, we focus on the changes of the wavelength of the maximum emission (Fig. 1c) and 0.5–0.8  $\mu\text{m}$  spectral slope (Fig. 1d) observed at different heliocentric distances. To achieve this goal, we calculated the spectral radiance factor,  $I(\lambda)$  expressed in arbitrary units, derived from Eq. (2) by removing the factors depending on the dust grain distribution and properties:

$$I(\lambda, g) \propto S(\lambda)q(\lambda) \frac{p(g, \lambda)}{4\pi} \quad (3)$$

The phase functions show scattering characteristics depending on composition and grain sizes: as a general rule, transparent grains, like water ice, are characterized by a very intense and collimated forward-scattering peak and complex resonance peaks at intermediate phase angles. Opaque materials, like troilite, show an almost isotropic scattering function. Semi-transparent particles, like ice tholins, have an intermediate behaviour with a prevalence of backscattering response. Grain size plays a fundamental role in the scattering mechanism: in the Rayleigh regime (particle size much smaller than the wavelength), light is equally forward- and backscattered while in the geometric regime (particle size much larger than the wavelength), backscattering dominates.

Fixing composition and grain size, the corresponding phase functions at the 27 visible wavelengths allow the derivation of the spectral radiance according to Eq. (3).

While  $p(g, \lambda)$  has a smooth response in the Rayleigh and in the geometric scattering regimes, the simulations, in general, show more fluctuations occurring for certain combinations of composition, grain size and  $\theta$  angle when the grain sizes are comparable to the wavelength, for example, at about 1  $\mu\text{m}$ . To mitigate this effect, we computed simulations for  $25^\circ \leq g \leq 125^\circ$ , corresponding to scattering angles of  $55^\circ \leq \theta \leq 155^\circ$ , at steps of  $1^\circ$  and then averaged the angular response of  $p(g, \lambda)$  on  $4^\circ$  bins, similar to the VIRTIS FOV. Similarly, the  $p(g)$  was computed for single grain radius (0.10–0.99  $\mu\text{m}$  (90 bins, 0.01  $\mu\text{m}$  per bin), 1.0–9.9  $\mu\text{m}$  (90 bins, 0.1  $\mu\text{m}$  per bin) and 10–50  $\mu\text{m}$  (40 bins, 1  $\mu\text{m}$  per bin)) and then averaged for each of the three families. Following this method, the  $I(\lambda, g)$  (from Eq. (3)) was derived for each composition and

size distribution. The  $\lambda_{\text{max}}$  and, after conversion in  $I/F$ , the 0.5–0.8  $\mu\text{m}$  spectral slope were derived from the simulated  $I(\lambda, g)$ . The corresponding values are shown in Figs. 3, 4, respectively.

### Modelling dust grain composition and size distribution in the coma

To disentangle the viewing and illumination geometry effects from particles' physical properties and to provide a more quantitative assessment of the changes observed in the  $\lambda_{\text{max}}$  time series, the light scattering on grains of different composition and radius was simulated. Optical and in situ measurements of the dust grain distribution in the 67P coma by Rosetta's instruments have revealed that the particles lie mainly within the 0.1  $\mu\text{m}$ –1 mm diameter range<sup>52</sup>. For particles smaller than 1 mm, the dust size distribution follows a power-law distribution with a coefficient varying between  $-2$  for heliocentric distances beyond 2 AU and  $-3.7$  at perihelion. Conversely, large grains (diameter greater than 1 mm) show a much steeper distribution with a coefficient of  $-4$  (ref. <sup>53</sup>). As small grains are the predominant population, in our analysis we preferentially model the spectral behaviour of small (radius less than 50  $\mu\text{m}$ ), compact grains. A similar grain size range has been also used by refs. <sup>28,37</sup> to model dust optical properties at infrared wavelengths and by ref. <sup>20</sup> to study thermal evolution and depletion of ice in dust aggregates along jets. The VIRTIS data are interpreted through the Mie theory, which can simulate the scattering properties of spherical grains of radius  $R$  and homogeneous composition having  $x = (2\pi R/\lambda) < 1,000$ . Such a limit corresponds to a maximum grain radius  $R = 50 \mu\text{m}$  in the visible spectral range (wavelength  $0.35 \leq \lambda \leq 1.0 \mu\text{m}$ ). To simulate the spectral response of larger grains, including fluffy aggregates<sup>7,54</sup>, it is necessary to use a geometric optics approximation<sup>55</sup>, which is beyond the scope of this work.

The changes occurring in the spectral radiance scattered by spherical grains of different composition and grain size distribution as a function of the phase angle ( $g$ ) are calculated for homogeneous particles made of possible cometary material endmembers, including water ice, organic ice tholin, troilite, amorphous carbon and magnesium silicate. For a given composition and phase angle, the single particle phase function  $p(g, \lambda)$  modulates the intensity of the scattered spectral radiance.

The theoretical trends of  $\lambda_{\text{max}}$  as a function of the scattering angle  $\theta = 180^\circ - g$  for particles of different composition and three grain size distributions (radius 0.10–0.99  $\mu\text{m}$ , 1.0–9.9  $\mu\text{m}$  and 10–50  $\mu\text{m}$ ) considered in this study, are shown in Fig. 3. In the same figure, we show the distribution of  $\lambda_{\text{max}}$  as measured from VIRTIS observations (Fig. 3f) selected during six MTP periods encompassing pre-perihelion (MTP015 and MTP016), perihelion (MTP020 and MTP021) and post-perihelion (MTP027 and MTP028) orbital phases. These periods are chosen because they offer the widest spread in scattering angle necessary to discriminate among different compositions and grain size distributions. Our analysis shows that, with the exception of transparent grains, for example, water ice and organic ice tholin, the response of opaque materials like amorphous carbon (Fig. 3a) and troilite (Fig. 3d) are characterized by a constant  $\lambda_{\text{max}}$  at about 0.5  $\mu\text{m}$  for the intermediate (1.0–9.9  $\mu\text{m}$ ) and large (10–50  $\mu\text{m}$ ) grain size distributions across the  $60^\circ \leq \theta \leq 150^\circ$  scattering angle range explored by VIRTIS. These grains are therefore not matching the observed distribution of the  $\lambda_{\text{max}}$ . Conversely, opaque-material submicrometre grains show a significant shift of the  $\lambda_{\text{max}}$  towards the red range, especially for  $\theta \leq 120^\circ$ .

Submicrometre grains made of magnesium silicate (Fig. 3b) show a linear increasing trend of  $\lambda_{\text{max}}$  from 0.56  $\mu\text{m}$  at  $\theta = 60^\circ$  to 0.63  $\mu\text{m}$  at  $\theta = 150^\circ$ , which is not compatible with VIRTIS data. In contrast, the  $\lambda_{\text{max}}$  of grains with radii larger than 1  $\mu\text{m}$  is dispersed around 0.55  $\mu\text{m}$ , a value too red to reproduce VIRTIS data at  $\theta > 100^\circ$  and too blue for  $\theta < 100^\circ$ .

Water-ice grains (Fig. 3e) offer the only solution matching VIRTIS data taken during the pre- (MTP015 and MTP016) and post-perihelion periods (MTP027 and MTP028) at scattering angles  $\theta > 100^\circ$ . During these phases, VIRTIS data are dispersed at  $\lambda_{\text{max}} < 0.5 \mu\text{m}$ , a behaviour

compatible with only the three water-ice grain-size distributions we have simulated.

Apart water ice, organic material grains (Fig. 3c) are the second end-member showing spectral similarities compatible with 67P particles: the 1–10  $\mu\text{m}$  and 10–50  $\mu\text{m}$  grain-size responses are in fact characterized by a linear decrease of  $\lambda_{\text{max}}$  from  $\theta = 60^\circ$  to  $120^\circ$ , making them compatible with observations acquired during perihelion and post-perihelion periods. In contrast, submicrometre grains show an extreme red colour across the entire range of scattering angles: having  $\lambda_{\text{max}} \geq 0.65 \mu\text{m}$ , they are marginally compatible only with VIRTIS observations taken at low  $\theta$  angles. If one assumes that during one month, for example, during the duration of one MTP period, the dominant composition and grain size is preserved, then the VIRTIS measurements taken during MTP028 at  $\theta > 130^\circ$  could be compatible not only with water ice but also with large (10–50  $\mu\text{m}$ ) organic material grains. As shown in Fig. 3a, amorphous carbon submicrometre-size (0.1–0.9  $\mu\text{m}$ ) grains are characterized by a reddening of  $\lambda_{\text{max}}$  for  $\theta < 90^\circ$ , similar to VIRTIS but with absolute values too low to exactly match the observations. Instead, the match is compatible for  $90^\circ < \theta < 120^\circ$ , making submicrometre amorphous carbon grains a good interloper between water-ice and organic grains. Finally, magnesium silicate (Fig. 3b) and troilite (Fig. 3d) submicrometre grains show  $\lambda_{\text{max}}$  trends at low scattering angles not compatible with VIRTIS data.

Composition and grain size distribution of the dust particles were further analysed through the 0.5–0.8  $\mu\text{m}$  spectral slope (Fig. 4) derived from VIRTIS data, which shows a peaked distribution at  $\theta = 100$ – $110^\circ$  (Fig. 4f) occurring in perihelion (MTP020 and MTP021) and post perihelion (MTP028) sequences. In our simulations, a similar behaviour is compatible with submicrometre-size grains made of organic matter (black points in Fig. 4e) or with greater than 1  $\mu\text{m}$  water-ice grains (blue and red points in Fig. 4d). The absolute value of the slope on the peak (about  $3 \mu\text{m}^{-1}$ ) measured at perihelion (MTP020, cyan points) is similar to theoretical values for organic matter. Being transparent, water-ice particles are characterized by a peculiar behaviour: submicrometre grains show a blue colour response (negative slope, black points in Fig. 4d) not compatible with any VIRTIS observation, whereas large grains (blue and red points) are neutral to moderately red and could contribute to the VIRTIS observed slope at any scattering angle. Opaque grains with sizes greater than 1  $\mu\text{m}$  show an almost constant neutral to moderately red ( $0.5 \mu\text{m}^{-1}$ ) slope within the entire scattering angle range (see red and blue points for amorphous carbon in Fig. 4a and troilite in Fig. 4b). We remark that while the  $\lambda_{\text{max}}$  trends of magnesium silicate (Fig. 3b) grains are not compatible with VIRTIS data, their spectral slopes values (Fig. 4c) for radii greater than 1  $\mu\text{m}$  are remarkably similar for measurements taken during the post-perihelion period.

Extended Data Table 3 shows the compatibility scheme of the  $\lambda_{\text{max}}$  and 0.5–0.8  $\mu\text{m}$  spectral slope values for the different composition endmembers, grain radius and scattering angle compared with the VIRTIS data.

## Data availability

The VIRTIS calibrated data are publicly available through the European Space Agency's Planetary Science Archive website (<https://archives.esac.esa.int/psa/>).

31. Vallat, C. et al. The science planning process on the Rosetta mission. *Acta Astronaut.* **133**, 244–257 (2017).
32. Filacchione, G. et al. On-ground characterization of Rosetta/VIRTIS-M. II. Spatial and radiometric calibrations. *Rev. Sci. Instrum.* **77**, 103106–103106-9 (2006).

33. Acton, C. H. Ancillary data services of NASA's Navigation and Ancillary Information Facility. *Planet. Space Sci.* **44**, 65–70 (1996).
34. Jorda, L. et al. The global shape, density and rotation of comet 67P/Churyumov-Gerasimenko from preperihelion Rosetta/OSIRIS observations. *Icarus* **277**, 257–278 (2016).
35. Filacchione, G. Calibrazioni a terra e prestazioni in volo di spettrometri ad immagine nel visibile e nel vicino infrarosso per l'esplorazione planetaria. PhD thesis, Università di Napoli Federico (2006).
36. Bohlin, R. C. & Gilliland, R. L. Hubble Space Telescope absolute spectrophotometry of Vega from the far-ultraviolet to the infrared. *Astron. J.* **127**, 3508–3515 (2004).
37. Rinaldi, G. et al. Summer outbursts in the coma of comet 67P/Churyumov-Gerasimenko as observed by Rosetta-VIRTIS. *Mon. Not. R. Astron. Soc.* **481**, 1235–1250 (2018).
38. Ciarniello, M. et al. Photometric properties of comet 67P/Churyumov-Gerasimenko from VIRTIS-M onboard Rosetta. *Astron. Astrophys.* **583**, A31 (2015).
39. Capaccioni, F. et al. The organic-rich surface of comet 67P/Churyumov-Gerasimenko as seen by VIRTIS/Rosetta. *Science* **347**, aaa0628 (2015).
40. Raponi, A. et al. The temporal evolution of exposed water ice-rich areas on the surface of 67P/Churyumov-Gerasimenko: spectral analysis. *Mon. Not. R. Astron. Soc.* **462**, S476–S490 (2016).
41. De Sanctis, M. C. et al. The diurnal cycle of water ice on comet 67P/Churyumov-Gerasimenko. *Nature* **525**, 500–503 (2015).
42. Filacchione, G. et al. Seasonal exposure of carbon dioxide ice on the nucleus of comet 67P/Churyumov-Gerasimenko. *Science* **354**, 1563–1566 (2016).
43. Hilchenbach, M. et al. Comet 67P/Churyumov-Gerasimenko: close-up on dust particle fragments. *Astrophys. J. Lett.* **816**, L32 (2016).
44. Bohren, C. F. & Huffman, D. R. *Absorption and Scattering of Light by Small Particles* (Wiley, 1983).
45. Pollack, J. B. et al. Composition and radiative properties of grains in molecular clouds and accretion disks. *Astrophys. J.* **421**, 615–639 (1994).
46. Cuzzi, J. N., Estrada, P. R. & Sanford, D. S. Utilitarian opacity model for aggregate particles in protoplanetary nebulae and exoplanet atmospheres. *Astrophys. J. Suppl. Ser.* **210**, 21 (2014).
47. Zubko, V. G. et al. Optical constants of cosmic carbon analogue grains—I. Simulation of clustering by a modified continuous distribution of ellipsoids. *Mon. Not. R. Astron. Soc.* **282**, 1321–1329 (1996).
48. Dorschner, J. et al. Steps towards interstellar silicate mineralogy. II. Study of Mg-Fe-silicate glasses of variable composition. *Astron. Astrophys.* **300**, 503–520 (1995).
49. Filacchione, G. et al. Comet 67P/CG nucleus composition and comparison to other comets. *Space Sci. Rev.* **215**, 19 (2019).
50. Fink, U. & Rinaldi, G. Coma dust scattering concepts applied to the Rosetta mission. *Icarus* **257**, 9–22 (2015).
51. Kurucz, R. L. Synthetic infrared spectra. In *Proc. IAU Symposium 154, Infrared Solar Physics* (eds Rabin, D. M. et al.) 523 (Kluwer, 1994).
52. Rotundi, A. Dust measurements in the coma of comet 67P/Churyumov-Gerasimenko inbound to the Sun. *Science* **347**, aaa3905 (2015).
53. Fulle, M. et al. Evolution of the dust size distribution of comet 67P/Churyumov-Gerasimenko from 2.2 au to perihelion. *Astrophys. J.* **821**, 19 (2016).
54. Fulle, M. et al. Density and charge of pristine fluffy particles from comet 67P/Churyumov-Gerasimenko. *Astrophys. J. Lett.* **802**, L12 (2015).
55. Grynko, Y. & Shkuratov, Y. G. in *Light Scattering Reviews 3* (ed. Kokhanovsky, A. A.) Ch. 9 (Springer Praxis, 2008).

**Acknowledgements** The authors thank the following institutions and agencies that supported this work: Italian Space Agency (ASI-Italy), Centre National d'Etudes Spatiales (CNES-France), Deutsches Zentrum für Luft-und Raumfahrt (DLR-Germany). VIRTIS was built by a consortium from Italy, France and Germany, under the scientific responsibility of IAPS, Istituto di Astrofisica e Planetologia Spaziali of INAF, Rome, which lead also the scientific operations. The VIRTIS instrument development for ESA has been funded and managed by ASI (Italy), with contributions from Observatoire de Meudon (France) financed by CNES and from DLR (Germany). The VIRTIS instrument industrial prime contractor was former Officine Galileo, now Leonardo Company, in Campi Bisenzio, Florence, Italy. This research has made use of NASA's Astrophysics Data System.

**Author contributions** The paper is a collective effort by the VIRTIS dust working group. F.G. as the main author of the paper calibrated, processed and interpreted VIRTIS data. F.C. as the VIRTIS principal investigator managed the experiment; M.C., A.R. and G.R. supported the Mie scattering calculations and data processing; F.C. and D.B.-M. planned 67P coma observations by VIRTIS; F.C. and G.F. planned the 67P nucleus observations by VIRTIS; S.E. provided geometry files for coma and nucleus observations; all authors, including M.C.D.C., G.A., V.M., M.F., A.L. and S.M., have contributed to the discussion of the results.

**Competing interests** The authors declare no competing interests.

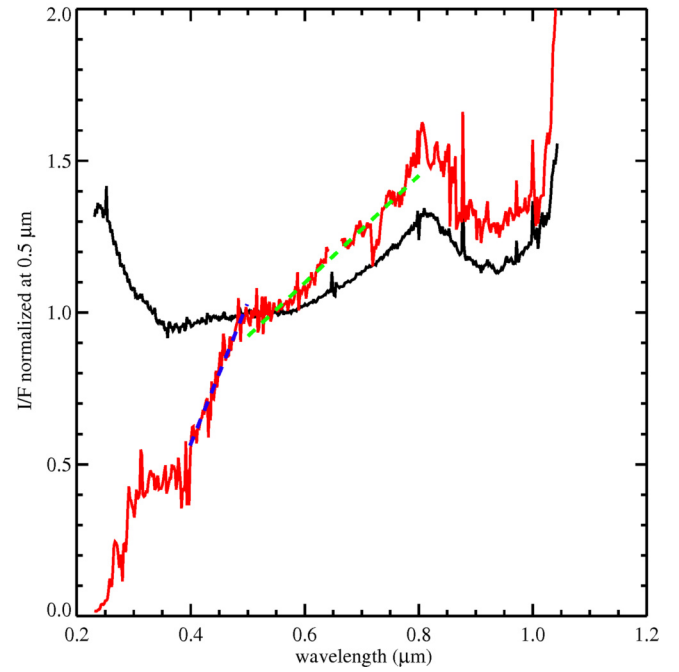
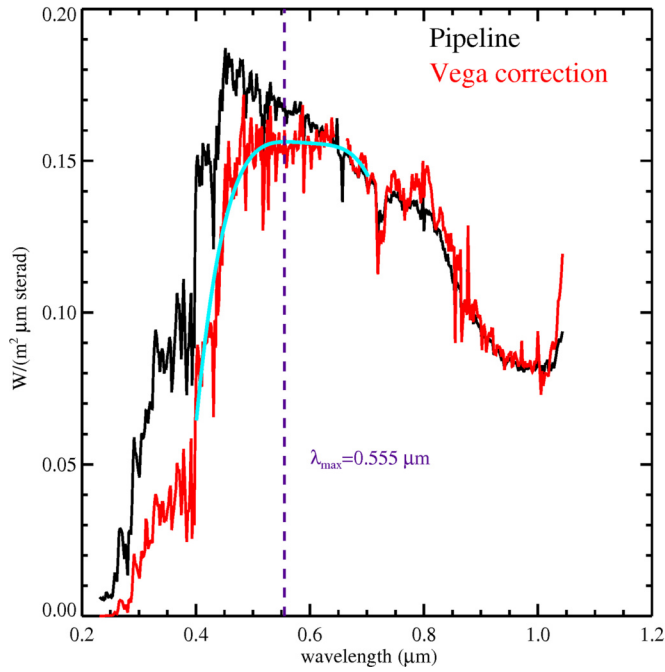
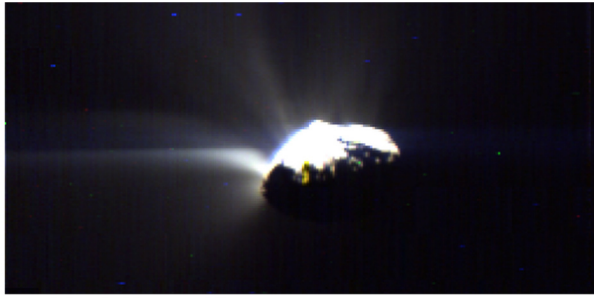
## Additional information

**Correspondence and requests for materials** should be addressed to G.F.

**Peer review information** Nature thanks Evgenij S. Zubko and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

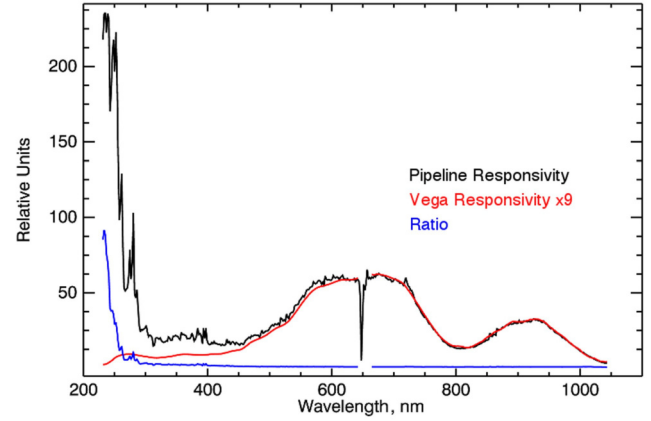
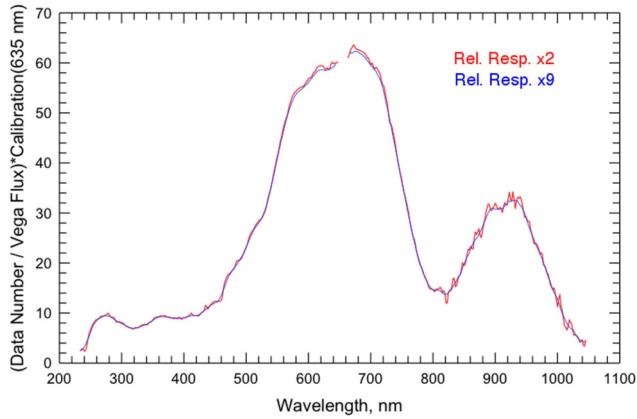
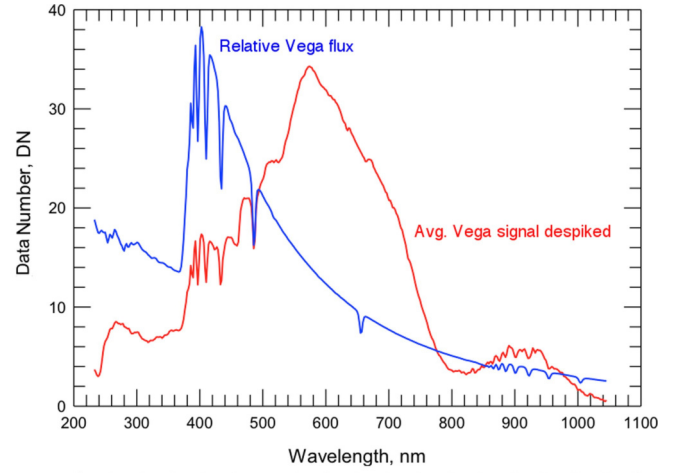
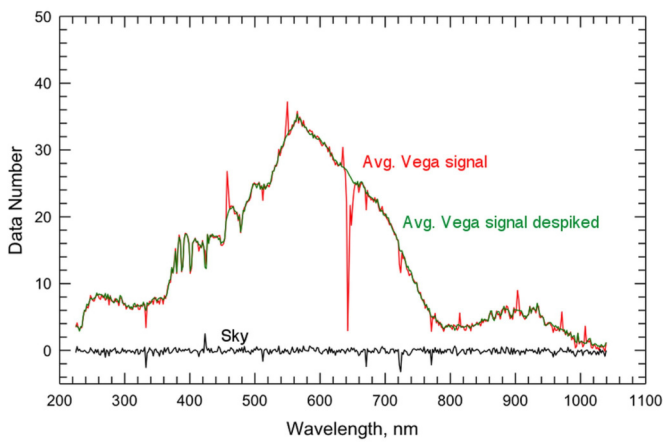
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





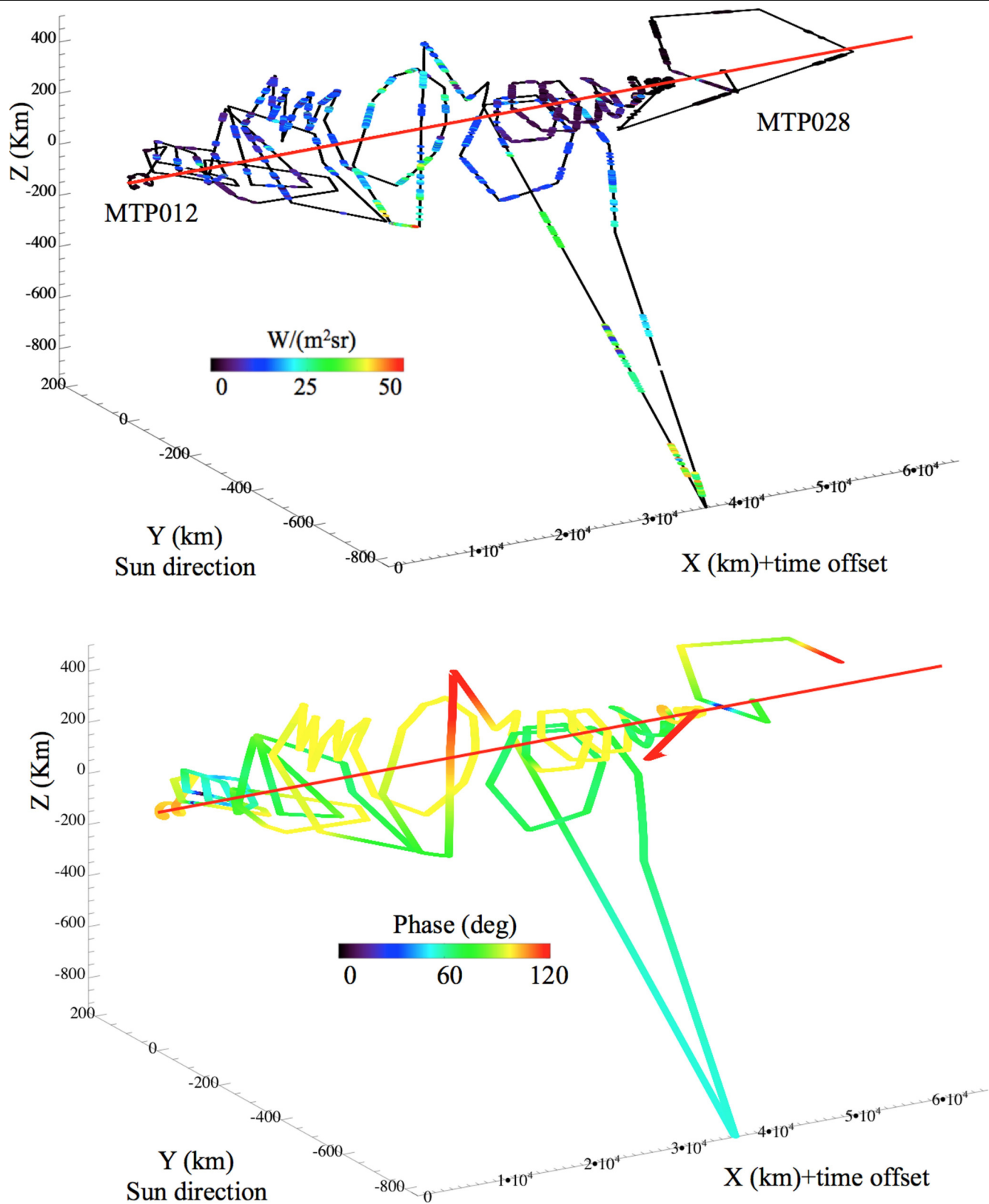
**Extended Data Fig. 1 | Example of a typical VIRTIS-M observation of 67P nucleus and coma.** Top left: visible colours RGB = (0.7, 0.55, 0.44  $\mu m$ ) image, stretched to saturate the nucleus and enhance the visibility of jets in the coma. Top right: tangent altitude image where the green mask corresponds to the annulus containing all pixels with a tangent altitude between 1 and 2.5 km from the limb. Bottom left: average radiance spectra as derived from official

pipeline (black curve) and after correction with Vega data (red curve). The fourth-degree polynomial fit to the corrected radiance is shown (cyan curve). The retrieved maximum emission wavelength on the fit is indicated by the magenta dashed line. Bottom right: corresponding  $I/F$  spectra normalized at  $0.5 \mu m$ . The best-fitting slopes in the  $0.4-0.5$  and  $0.5-0.8 \mu m$  ranges are indicated by blue and green dashed lines, respectively.



**Extended Data Fig. 2 | Vega star signal and derived VIRTIS responsivity.** Top left: average spectrum of Vega, in DN, from observation V1\_00402035638.QUB (red curve) and spectrum corrected for dark current, despiked and filter notch removed (green curve). Note the correlation of negative spikes in the Vega and average sky spectrum (black curve). Owing to the instrumental point spread function and spectral tilt, the signal is an average taken from 18 pixels. Top right: average spectrum of Vega derived from the four observations listed in Extended Data Table 2 after having applied a processing similar to the one

shown in the previous plot. The curve is averaged with a two-point running boxcar filter. The Vega flux<sup>36</sup> is shown in relative units (blue curve). Bottom left: VIRTIS responsivity derived from Vega signal averaged with a two-point running boxcar filter (red curve) and nine points (blue curve) after normalization at 0.635  $\mu\text{m}$  above the standard responsivity value. Bottom right: comparison between standard pipeline (black curve) and Vega responsivities with a nine-point running boxcar filter (red curve). The ratio between the two responses is the blue curve.



**Extended Data Fig. 3 | Rosetta spacecraft three-dimensional trajectory and solar phase angle variations with time.** Top: Rosetta trajectory in the 67P XYZ reference frame. Points along the X axis are shown starting from Rosetta's position at 2015-01-13T23:28:53 (MTP012) with an increment of 1 km every 20 min to improve visualization. The Y axis is oriented towards the Sun and the

Z axis is perpendicular to the orbital plane. The red line indicates the position of the nucleus along the X axis. The integrated radiance as measured on each observation is reported according to the colour scale. Bottom: variation of the solar phase angle (Sun–nucleus centre–Rosetta) during the mission.



## Article

Extended Data Table 1 | Rosetta's calendar MTP periods dates and heliocentric distance

Period	Start Time	End Time	Heliocentric distance (AU)
MTP012	2015-01-13T23:28:53	2015-02-10T23:23:53	2.552-2.344
MTP013	2015-02-10T23:23:53	2015-03-10T23:23:53	2.344-2.130
MTP014	2015-03-10T23:23:53	2015-04-08T23:23:53	2.130-1.967
MTP015	2015-04-08T23:23:53	2015-05-05T23:23:53	1.967-1.703
MTP016	2015-05-05T23:23:53	2015-06-02T23:23:53	1.703-1.508
MTP017	2015-06-02T23:23:53	2015-06-30T23:23:53	1.508-1.350
MTP018	2015-06-30T23:23:53	2015-07-28T23:23:52	1.350-1.257
MTP019	2015-07-28T23:23:52	2015-08-25T23:23:52	1.257-1.253
MTP020	2015-08-25T23:23:52	2015-09-22T23:23:52	1.253-1.340
MTP021	2015-09-22T23:23:52	2015-10-20T23:23:52	1.340-1.494
MTP022	2015-10-20T23:23:52	2015-11-17T23:23:52	1.494-1.686
MTP023	2015-11-17T23:23:52	2015-12-15T23:28:52	1.686-1.897
MTP024	2015-12-15T23:28:52	2016-01-12T23:28:52	1.897-2.113
MTP025	2016-01-12T23:28:52	2016-02-09T23:28:52	2.113-2.327
MTP026	2016-02-09T23:28:52	2016-03-08T23:28:52	2.327-2.536
MTP027	2016-03-08T23:28:52	2016-04-05T23:28:52	2.536-2.738
MTP028	2016-04-05T23:28:52	2016-05-03T23:27:51	2.738-2.933

Extended Data Table 2 | List of VIRTIS observations of the Vega star

Observation	Start Time	End Time	(Band, Sample, Line)	Int. Time (s)
V1_00402035638.QUB	2015-09-28T04:35:17.781	2015-09-28T05:04:11.405	(432,256,29)	50
V1_00403369562.QUB	2015-10-13T15:07:22.178	2015-10-13T15:38:15.817	(432,256,31)	50
V1_00403848118.QUB	2015-10-19T04:03:18.373	2015-10-19T04:34:12.013	(432,256,31)	50
V1_00406049718.QUB	2015-11-13T15:36:39.065	2015-11-13T16:03:32.757	(432,256,27)	50

Extended Data Table 3 | Summary of spectral indicator compatibility

Composition	Grain radius ( $\mu\text{m}$ )	Scattering Angle (deg)									
		60	70	80	90	100	110	120	130	140	150
Amorphous Carbon	0.10-0.99										
	1.0-9.9										
	10-50										
Mg-silicate	0.10-0.99										
	1.0-9.9										
	10-50										
Organics	0.10-0.99										
	1.0-9.9										
	10-50										
Troilite	0.10-0.99										
	1.0-9.9										
	10-50										
Water Ice	0.10-0.99										
	1.0-9.9										
	10-50										

Compatible (<1 stdev)

Marginal compatible (<2 stdev)

Not compatible (>2 stdev)

N/A coverage

Pre-Perihelion

Perihelion

Post-Perihelion

0.5-0.8  $\mu\text{m}$  spectral slope

$\lambda_{\text{MAX}}$

The table shows the compatibility of spectral indicators ( $\lambda_{\text{MAX}}$ , 0.5–0.8  $\mu\text{m}$  spectral slopes) between VIRTIS observations and Mie scattering simulations for a given composition, grain radius range and scattering angle. The compatibility between observations and scattering simulations is colour coded according to the key shown on the top right. Each cell is divided in four fields showing  $\lambda_{\text{MAX}}$  and to 0.5–0.8  $\mu\text{m}$  slope values during pre-perihelion, perihelion and post-perihelion epochs (bottom right key).



# Demonstration of cooling by the Muon Ionization Cooling Experiment

<https://doi.org/10.1038/s41586-020-1958-9>

MICE collaboration\*

Received: 22 July 2019

Accepted: 13 December 2019

Published online: 5 February 2020

Open access

The use of accelerated beams of electrons, protons or ions has furthered the development of nearly every scientific discipline. However, high-energy muon beams of equivalent quality have not yet been delivered. Muon beams can be created through the decay of pions produced by the interaction of a proton beam with a target. Such ‘tertiary’ beams have much lower brightness than those created by accelerating electrons, protons or ions. High-brightness muon beams comparable to those produced by state-of-the-art electron, proton and ion accelerators could facilitate the study of lepton–antilepton collisions at extremely high energies and provide well characterized neutrino beams<sup>1–6</sup>. Such muon beams could be realized using ionization cooling, which has been proposed to increase muon-beam brightness<sup>7,8</sup>. Here we report the realization of ionization cooling, which was confirmed by the observation of an increased number of low-amplitude muons after passage of the muon beam through an absorber, as well as an increase in the corresponding phase-space density. The simulated performance of the ionization cooling system is consistent with the measured data, validating designs of the ionization cooling channel in which the cooling process is repeated to produce a substantial cooling effect<sup>9–11</sup>. The results presented here are an important step towards achieving the muon-beam quality required to search for phenomena at energy scales beyond the reach of the Large Hadron Collider at a facility of equivalent or reduced footprint<sup>6</sup>.

## High-quality muon beams

Fundamental insights into the structure of matter and the nature of its elementary constituents have been obtained using beams of charged particles. The use of time-varying electromagnetic fields to produce sustained acceleration was pioneered in the 1930s<sup>12–14</sup>. Since then, high-energy and high-brightness particle accelerators have delivered electron, proton and ion beams for applications ranging from the search for new phenomena in the interactions of quarks and leptons to the study of nuclear physics, materials science and biology.

Muon beams can be created using a proton beam striking a target to produce a secondary beam comprising many particle species including pions, kaons and muons. The pions and kaons decay to produce additional muons, which are captured by electromagnetic beamline elements to produce a tertiary muon beam. Capture must be realized on a timescale compatible with the muon lifetime at rest, 2.2  $\mu$ s. Without acceleration, the energy and intensity of the muon beam is limited by the energy and intensity of the primary proton beam and the efficiency with which muons are captured.

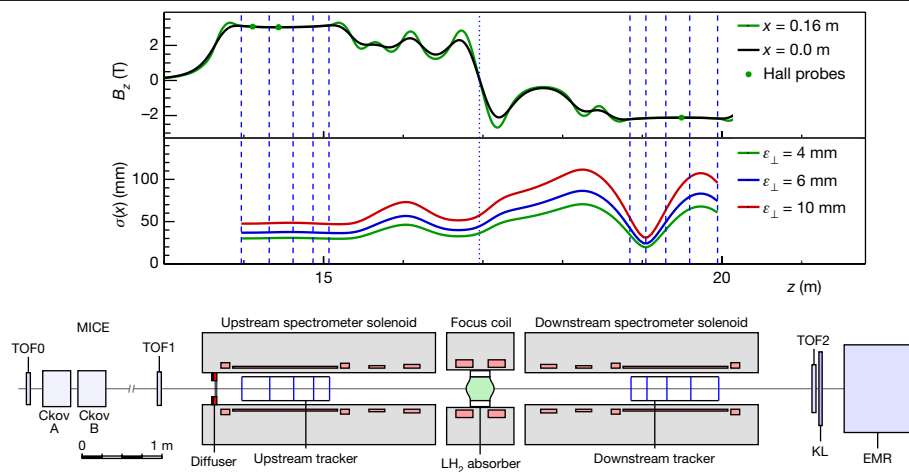
Accelerated high-brightness muon beams have been proposed as a source of neutrinos at neutrino factories and for the delivery of multi-TeV lepton–antilepton collisions at muon colliders<sup>1–6</sup>. Muons have attractive properties for the delivery of high-energy collisions. The muon is a fundamental particle with mass 207 times that of the electron. This high mass results in suppression of synchrotron radiation, potentially enabling collisions between beams of muons and

antimuons at energies far in excess of those that can be achieved in an electron–positron collider, such as the proposed International Linear Collider<sup>15</sup>, the Compact Linear Collider<sup>16</sup>, the Circular Electron–Positron Collider<sup>17</sup> and the electron–positron option of the Future Circular Collider<sup>18</sup>. The virtual absence of synchrotron radiation makes it possible to build a substantially smaller facility with the same or greater physics reach.

The energy available in collisions between the constituent gluons and quarks in proton–proton collisions is considerably less than the energy of the proton beam because the colliding quarks and gluons each carry only a fraction of the proton’s momentum. Muons carry the full energy of the beam, making muon colliders attractive for the study of particle physics beyond the energy reach of facilities such as the Large Hadron Collider<sup>19</sup>.

Most of the proposals for accelerated muon beams exploit the proton-driven muon-beam production scheme outlined above and use beam cooling to increase the brightness of the tertiary muon beam before acceleration and storage to ensure sufficient luminosity or beam current. Four cooling techniques are in use at particle accelerators: synchrotron radiation cooling<sup>20</sup>, laser cooling<sup>21</sup>, stochastic cooling<sup>22</sup> and electron cooling<sup>23</sup>. In each case, the time required to cool the beam is long compared to the muon lifetime. Frictional cooling of muons, in which muons are electrostatically accelerated through an energy-absorbing medium at energies significantly below 1 MeV, has been demonstrated but with low efficiency<sup>24–26</sup>.

\*A list of participants and their affiliations appears at the end of the paper.



**Fig. 1 | The MICE apparatus, the calculated magnetic field and the nominal horizontal width of the beam.** The modelled field,  $B_z$ , is shown on the beam axis (black line) and at 160 mm from the axis (green line) in the horizontal plane. The readings of Hall probes situated at 160 mm from the beam axis are also shown. Vertical lines indicate the positions of the tracker stations (dashed

lines) and the absorber (dotted line). The nominal r.m.s. beam width,  $\sigma(x)$ , is calculated assuming a nominal input beam and using linear beam transport equations. See text for the description of the MICE apparatus. TOF0, TOF1 and TOF2 are time-of-flight detector stations; KL is a lead–scintillator pre-shower detector; EMR is the Electron–Muon Ranger.

The technique demonstrated in this study, ionization cooling<sup>7,8</sup>, is based on a suitably prepared beam passing through an appropriate material (the absorber) and losing momentum through ionization. Radio-frequency cavities restore momentum only along the beam direction. Passing the muon beam through a repeating lattice of material and accelerators causes the ionization cooling effect to build up in a time much shorter than the muon lifetime<sup>9–11</sup>. Acceleration of a muon beam in a radio-frequency accelerator has recently been demonstrated<sup>27</sup>, and reduced beam heating, damped by the ionization cooling effect, has been observed<sup>28</sup>. Ionization cooling has not been demonstrated so far. Experimental validation of the technique is important for the development of muon accelerators. The international Muon Ionization Cooling Experiment (MICE; <http://mice.iit.edu>) was designed to demonstrate transverse ionization cooling, the realization of which is presented here.

The brightness of a particle beam can be characterized by the number of particles in the beam and the volume occupied by the beam in position–momentum phase space. The phase-space volume occupied by the beam and the phase-space density of the beam are conserved quantities in a conventional accelerator without cooling. The phase space considered here is the position and momentum transverse to the direction of travel of the beam,  $\mathbf{u} = (x, p_x, y, p_y)$ , where  $x$  and  $y$  are coordinates perpendicular to the beam line, and  $p_x$  and  $p_y$  are the corresponding components of the momentum. The  $z$  axis is the nominal beam axis.

The normalized root-mean-square (r.m.s.) emittance is conventionally used as an indicator of the phase-space volume occupied by the beam<sup>29</sup>, but this quantity is not conserved when scraping or optical aberrations affect the edge of the beam. The distribution of amplitudes<sup>30,31</sup> is used here to study effects in the core of the beam. The amplitude of a particle is the distance of the particle from the beam centroid in normalized phase space, and is a conserved quantity in a conventional accelerator without cooling. The phase-space density of the beam is also directly studied using a  $k$ -nearest-neighbour technique<sup>32</sup>.

## MICE cooling apparatus

The MICE collaboration has built a tightly focusing solenoid lattice, absorbers and instrumentation to demonstrate the ionization cooling of muons. A schematic of the apparatus is shown in Fig. 1.

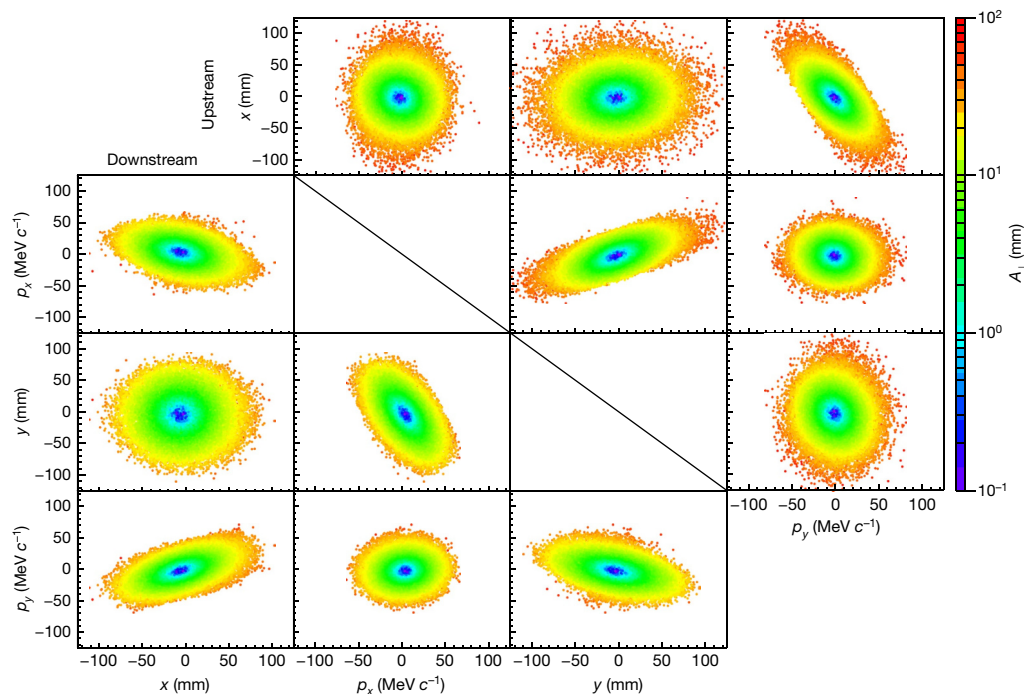
A transfer line<sup>33–35</sup> brought a beam, composed mostly of muons, from a target<sup>36</sup> in the ISIS synchrotron<sup>37</sup> to the cooling apparatus. The central momentum of the muons could be tuned between  $140 \text{ MeV } c^{-1}$  and  $240 \text{ MeV } c^{-1}$  ( $c$ , speed of light in vacuum). A variable-thickness brass and tungsten diffuser allowed the emittance of the incident beam to be varied between 4 mm and 10 mm.

The tight focusing (low  $\beta$  function) and large acceptance required by the cooling section was achieved using 12 superconducting solenoids. The solenoids were contained in three warm-bore modules cooled by closed-cycle cryocoolers. The upstream and downstream modules (spectrometer solenoids) were identical, each containing three coils to provide a uniform field region of up to 4 T within the 400-mm-diameter warm bore for momentum measurement, as well as two ‘matching’ coils to match the beam to the central pair of closely spaced ‘focus’ coils, which focus the beam onto the absorber. The focus coils were designed to enable peak on-axis fields of up to 3.5 T within one module with a 500-mm-diameter warm bore containing the absorbers.

For the experiment reported here the focus coils were operated in ‘flip’ mode with a field reversal at the centre. Because the magnetic lattice was tightly coupled, the cold mass-suspension systems of the modules were designed to withstand longitudinal cold-to-warm forces of several hundred kN, which could arise during an unbalanced quench of the system. At maximum field, the inter-coil force on the focus coil cold mass was of the order of 2 MN. The total energy stored in the magnetic system was of the order of 5 MJ and the system was protected by both active and passive quench-protection systems. The normal charging and discharging time of the solenoids was several hours. The entire magnetic channel was partially enclosed by a 150-mm-thick soft-iron return yoke for external magnetic shielding. The magnetic fields in the tracking volumes were monitored during operation using calibrated Hall probes.

One of the matching coils in the downstream spectrometer solenoid was not operable owing to a failure of a superconducting lead. Although this necessitated a compromise in the lattice optics and acceptance, the flexibility of the magnetic lattice was exploited to ensure a clear cooling measurement.

The amplitude acceptance of approximately 30 mm, above which particles scrape, was large compared to that of a typical accelerator. Even so, considerable scraping was expected and observed for the highest-emittance beams. Ionization cooling cells with even larger acceptances, producing less scraping, have been designed<sup>9–11</sup>.



**Fig. 2 | Beam distribution in phase space for the 6–140 Full LH2 setting of MICE.** Measured beam distribution in the upstream tracker (above the diagonal) and in the downstream tracker (below the diagonal). The measured coordinates of the particles are coloured according to the amplitude  $A_{\perp}$  of the particle.

The magnetic lattice of MICE, shown in Fig. 1, was tuned so that the focus of the beam was near the absorber, resulting in a small beam width and large angular divergence. The tight focusing, corresponding to a nominal transverse  $\beta$  function of around 430 mm at the centre of the absorber, yielded an optimal cooling performance.

Materials with low atomic number, such as lithium and hydrogen, have a long radiation length relative to the rate of energy loss, and consequently better cooling performance, making them ideal absorber materials. Therefore, cooling by both liquid-hydrogen and lithium hydride absorbers was studied.

The liquid hydrogen was contained within a 22-l vessel<sup>38</sup> in the warm bore of the focus coil. Hydrogen was liquefied by a cryocooler and piped through the focus coil module into the absorber body. When filled, the absorber presented  $349.6 \pm 0.2$  mm of liquid hydrogen along the beam axis with a density of  $0.07053 \pm 0.00008$  g cm<sup>-3</sup> (all uncertainties represent the standard error). The liquid hydrogen was contained between a pair of aluminium windows covered by multi-layer insulation. A second pair of windows provided a secondary barrier to protect against failure of the primary containment windows. These windows were designed to be as thin as possible so that any scattering in them would not cause substantial heating. The total thickness of all four windows on the beam axis was  $0.79 \pm 0.01$  mm.

The lithium hydride absorber was a disk of thickness  $65.37 \pm 0.02$  mm with a density of  $0.6957 \pm 0.0006$  g cm<sup>-3</sup>. The isotopic composition of the lithium used to produce the absorber was 95% <sup>6</sup>Li and 5% <sup>7</sup>Li. The cylinder had a thin coating of parylene to prevent ingress of water or oxygen. Configurations with the empty liquid-hydrogen containment vessel and with no absorber were also studied.

## MICE beam instrumentation

Detectors placed upstream and downstream of the apparatus measured the momentum, position and species of each particle entering and leaving the cooling channel in order to reconstruct the full four-dimensional phase space, including the angular momentum introduced by the solenoids. Particles were recorded by the apparatus

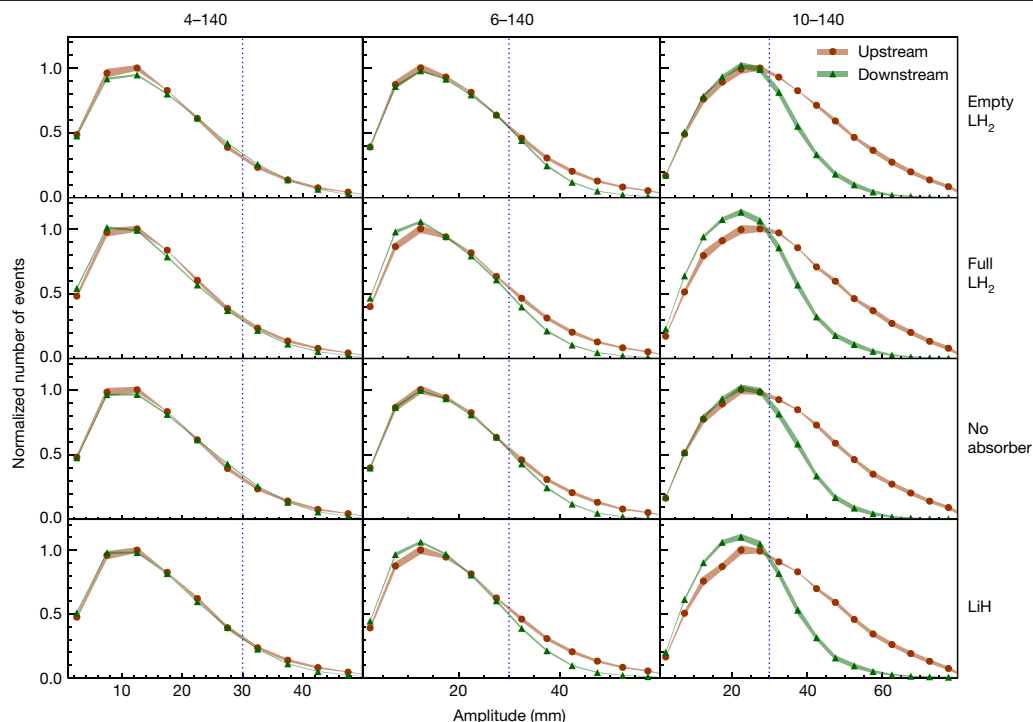
one at a time, which enabled high-precision instrumentation to be used and particles other than muons to be excluded from the analysis. Each ensemble of muons was accumulated over a number of hours. This is acceptable because space-charge effects are not expected at a neutrino factory and in a muon collider they become important only at very low longitudinal emittance<sup>39</sup>. Data-taking periods for each absorber were separated by a period of weeks owing to operational practicalities. The phase-space distribution of the resulting ensemble was reconstructed using the upstream and downstream detectors. The emittance reconstruction in the upstream detector system is described in ref. <sup>40</sup>.

Upstream of the cooling apparatus, two time-of-flight (TOF) detectors<sup>41</sup> measured the particle velocity. A complementary velocity measurement was made upstream by the threshold Cherenkov counters Ckov A and Ckov B<sup>42</sup>. Scintillating fibre trackers, positioned in the uniform-field region of each of the two spectrometer solenoids, measured the particle position and momentum upstream and downstream of the absorber<sup>43,44</sup>. Downstream, an additional TOF detector<sup>45</sup>, a mixed lead-scintillator pre-shower detector and a totally active scintillator calorimeter, the Electron–Muon Ranger<sup>46,47</sup>, identified electrons produced by muon decay and allowed cross-validation of the measurements made by the upstream detectors and the trackers.

Each tracker consisted of five planar scintillating-fibre stations. Each station comprised three views; each view was composed of two layers of 350- $\mu$ m-diameter scintillating fibres positioned at an angle of 120° with respect to the other views. The fibres were read out by cryogenic visible-light photon counters<sup>48</sup>. The position of a particle crossing the tracker was inferred from the coincidence of signals from the fibres, and the momentum was calculated by fitting a helical trajectory to the signal positions, with appropriate consideration for energy loss and scattering in the fibres.

Each TOF detector was constructed from two orthogonal planes of scintillator slabs. Photomultiplier tubes at each end of every TOF detector slab were used to determine the time at which a muon passed through the apparatus with a 60-ps resolution<sup>41</sup>. The momentum resolution of particles with a small helix radius in the tracker was improved





**Fig. 3 | Muon amplitudes measured by MICE.** The measured upstream distributions are shown by red circles while the downstream distributions are shown by green triangles. Both upstream and downstream distributions are normalized to the bin with the most entries in the upstream distribution (see text). Coloured bands show the estimated standard error, which is dominated

by systematic uncertainties. Vertical lines indicate the approximate channel acceptance above which scraping occurs. The number of events in each sample is listed in Extended Data Table 2. Data for each experimental configuration were accumulated in a single discrete period.

by combining the TOF measurement of velocity with the measurement of momentum in the tracker.

A detailed Monte Carlo simulation of the experiment was performed to study the resolution and efficiency of the instrumentation and to determine the expected performance of the cooling apparatus<sup>49,50</sup>. The simulation was found to give a good description of the data<sup>40</sup>.

### Demonstration of cooling

The data presented here were taken using beams with a nominal momentum of  $140 \text{ MeV } c^{-1}$  and a nominal normalized r.m.s. emittance in the upstream tracking volume of 4 mm, 6 mm and 10 mm; these settings are denoted as '4-140', '6-140' and '10-140', respectively. Beams with a higher emittance have more muons at high amplitudes and occupy a larger region in phase space. For each beam setting, two samples were considered for the analysis. The 'upstream sample' contained particles identified as muons by the upstream TOF detectors and tracker, for which the muon trajectory reconstructed in the upstream tracker was fully contained in the fiducial volume and for which the reconstructed momentum fell within the range  $135 \text{ MeV } c^{-1}$  to  $145 \text{ MeV } c^{-1}$  (which is considerably higher than the momentum resolution of the tracker,  $2 \text{ MeV } c^{-1}$ ). The 'downstream sample' was the subset of the upstream sample for which the reconstructed muons were fully contained in the fiducial volume of the downstream tracker. Each of the samples had between 30,000 and 170,000 events. Examples of the phase-space distributions of the particles in the two samples are shown in Fig. 2. The strong correlations between  $y$  and  $p_x$  and between  $x$  and  $p_y$  are due to the angular momentum introduced by the solenoidal field. The shorter tails along the semi-minor axis compared to the semi-major axis in these projections arise from scraping in the diffuser.

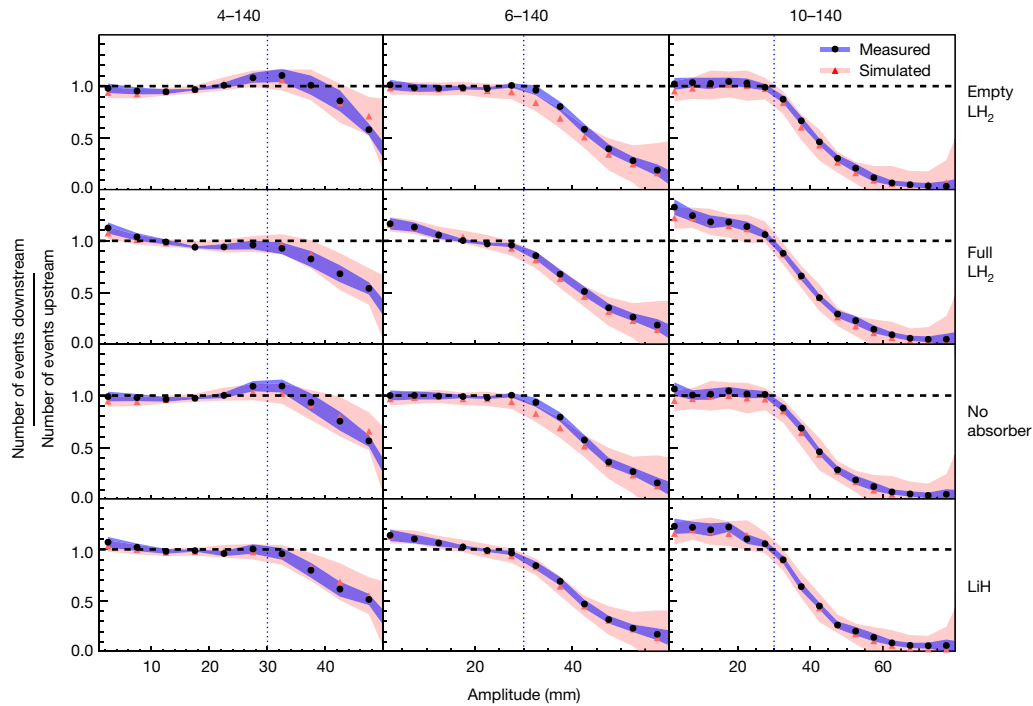
The distributions of amplitudes in the upstream and downstream samples for each of the 4-140, 6-140 and 10-140 datasets are shown in Fig. 3. The nominal acceptance of the magnetic channel is also

indicated. A correction has been made to account for the migration of events between amplitude bins due to the detector resolution and to account for inefficiency in the downstream detector system (see Methods). Distributions are shown for the measurements with an empty liquid-hydrogen vessel ('Empty LH<sub>2</sub>'), with a filled liquid-hydrogen vessel ('Full LH<sub>2</sub>'), with no absorber ('No absorber') and with the lithium hydride absorber ('LiH'). The distributions were normalized to allow a comparison of the shape of the distribution between different absorbers. Each pair of upstream and downstream amplitude distributions is scaled by  $1/N_{\text{max}}^u$ , where  $N_{\text{max}}^u$  is the number of events in the most populated bin in the upstream sample.

The behaviour of the beam at low amplitude is the key result of this study. For the 'No absorber' and 'Empty LH<sub>2</sub>' configurations, the number of events with low amplitude in the downstream sample is similar to that observed in the upstream sample. For the 6-140 and 10-140 configurations for both the 'Full LH<sub>2</sub>' and the 'LiH' samples, the number of events with low amplitude is considerably larger in the downstream sample than in the upstream sample. This indicates an increase in the number of particles in the beam core when an absorber is installed, which is expected if ionization cooling takes place. This effect can occur only because energy loss is a non-conservative process.

A reduction in the number of muons at high amplitude is also observed, especially for the 10-140 setting. Whereas part of this effect arises owing to migration of muons into the beam core, a substantial number of high-amplitude particles outside the beam acceptance intersected the beam pipe or fell outside the fiducial volume of the downstream tracker. The beam pipe was made of materials with higher atomic number than those of the absorber materials, so interactions in the beam pipe tended to be dominated by multiple Coulomb scattering, leading to beam loss.

A  $\chi^2$  test was performed to determine the confidence with which the null hypothesis that for the same input beam setting, the amplitude distributions in the downstream samples of the 'Full LH<sub>2</sub>' and 'Empty LH<sub>2</sub>'



**Fig. 4 | Downstream-to-upstream ratio of number of events in MICE.** A ratio greater than unity in the beam core, which is evidence of ionization cooling, is observed in the data obtained with the 6–140 and 10–140 beams with both the full  $\text{LH}_2$  absorber and the  $\text{LiH}$  absorber. The effect predicted from the simulation is shown in red and that measured is shown in black. The

corresponding shading shows the estimated standard error, which is dominated by systematic uncertainty. Vertical lines indicate the channel acceptance above which scraping occurs. The number of events in each sample is listed in Extended Data Table 2. Data for each experimental configuration were accumulated in a single discrete period.

configurations are compatible, and the amplitude distributions in the downstream samples of the ‘ $\text{LiH}$ ’ and ‘No absorber’ configurations are compatible. The test was performed on the uncorrected distributions using only statistical uncertainties. Systematic effects are the same for the pairs of distributions tested, and cancel. Assuming that this null hypothesis is correct, the probability of observing the effect seen in the data is considerably lower than  $10^{-5}$  for each beam setting and for each ‘Full  $\text{LH}_2$ ’–‘Empty  $\text{LH}_2$ ’ and ‘ $\text{LiH}$ ’–‘No absorber’ pair; therefore, the null hypothesis was rejected.

The fractional increase in the number of particles with low amplitude is most pronounced for the 10–140 beams. High-amplitude beams have high transverse emittance,  $\varepsilon_{\perp}$ , and a larger transverse momentum relative to the stochastic increase in transverse momentum due to scattering, so they undergo more cooling. For the magnet settings and beams studied here, heating due to multiple Coulomb scattering becomes dominant over ionization cooling at an emittance of around 4 mm. As a result, only modest cooling is observed for the 4–140 setting in both the ‘Full  $\text{LH}_2$ ’ and ‘ $\text{LiH}$ ’ configurations.

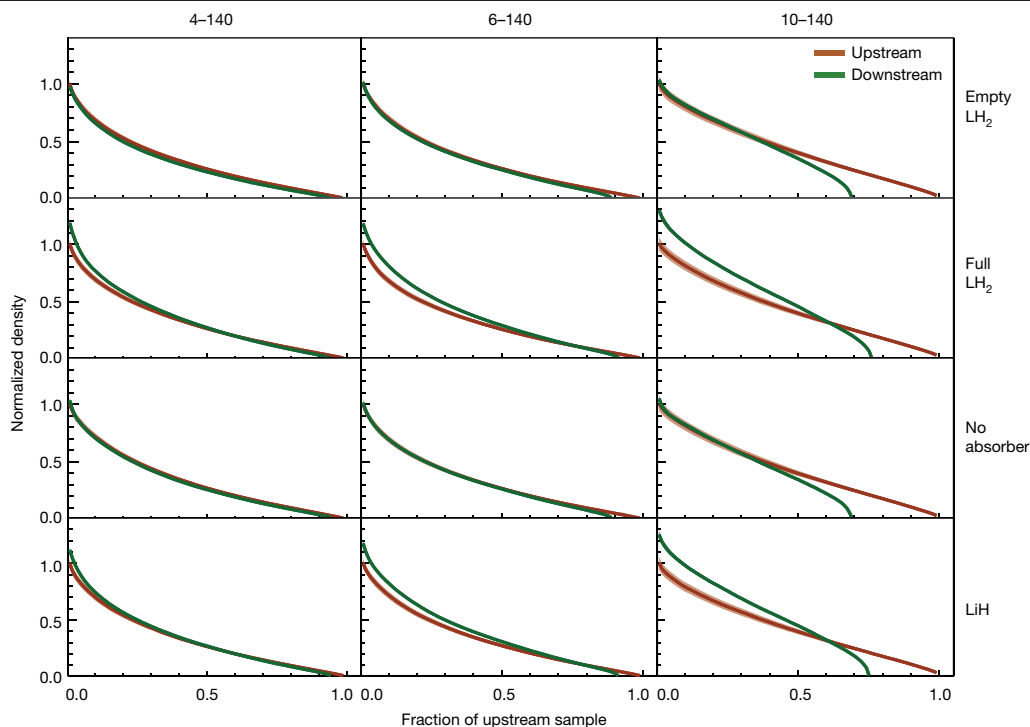
The ratios of the downstream to the upstream amplitude distributions are shown in Fig. 4. In the ‘No absorber’ and ‘Empty absorber’ configurations, the ratios are consistent with 1 for amplitudes of less than 30 mm, confirming the conservation of amplitude in this region, irrespective of the incident beam. Above 30 mm the ratios drop below unity, indicating that at high amplitude there are fewer muons downstream than upstream, as outlined above. The presence of the absorber windows does not strongly affect the amplitude distribution. For the 6–140 and 10–140 datasets, the addition of liquid-hydrogen or lithium hydride absorber material causes the ratios to rise above unity for the low-amplitude particles that correspond to the beam core. This indicates an increase in the number of particles in the beam core and demonstrates ionization cooling.

The density in phase space is an invariant of a symplectic system; therefore, an increase in phase-space density is also an unequivocal

demonstration of cooling. Figure 5 shows the normalized density of the upstream and downstream samples,  $\rho_i(\mathbf{u}_i)/\rho_0$ , as a function of  $\alpha$ , the fraction of the upstream sample that has a density greater than or equal to  $\rho_i$ . This is known as the quantile distribution. To enable comparison between different beam configurations, the densities for each configuration have been normalized to the peak density in the upstream tracker,  $\rho_0$ . To enable comparison between the upstream and downstream distributions, the fraction of the sample is always relative to the total number of events in the upstream sample. The transmission is the fraction of the beam for which the density in the downstream tracker reaches zero. For the ‘No absorber’ and ‘Empty  $\text{LH}_2$ ’ cases, the downstream density in the highest-density regions is indistinguishable from the upstream density. A small amount of scraping is observed for the 4–140 and 6–140 beams. More substantial scraping is observed for the 10–140 beam. In all cases, for ‘Full  $\text{LH}_2$ ’ and ‘ $\text{LiH}$ ’ the phase-space density increases, and the increase is greater for higher-emittance beams. These observations demonstrate the ionization cooling of the beam when an absorber is installed. In the presence of an absorber, beams with larger nominal emittance show a greater increase in density than those with a lower nominal emittance.

## Conclusions

Ionization cooling has been unequivocally demonstrated. We have built and operated a section of a solenoidal cooling channel and demonstrated the ionization cooling of muons using both liquid hydrogen and lithium hydride absorbers. The effect has been observed through the measurement of both an increase in the number of small-amplitude particles (Figs. 3, 4) and an increase in the phase-space density of the beam (Fig. 5). The results are well described by simulations (Fig. 4). This demonstration of ionization cooling is an important advance in the development of high-brightness muon beams. The seminal results presented in this paper encourage further development of high-brightness



**Fig. 5 | Normalized quantile distribution of the beam density in MICE.**

Upstream and downstream quantiles are indicated by orange and green lines, respectively, as a function of the fraction of the upstream sample. For each configuration, the density is normalized to the highest-density region in the

upstream sample. The estimated standard error is indicated by the thickness of the coloured bands and is dominated by systematic uncertainty. The number of events in each sample is listed in Extended Data Table 2. Data for each experimental configuration were accumulated in a single discrete period.

muon beams as a tool for the investigation of the fundamental properties of matter.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1958-9>.

- Neuffer, D. V. & Palmer, R. B. A high-energy high-luminosity  $\mu^+\mu^-$  collider. *AIP Conf. Proc.* **356**, 344–358 (1996).
- Geer, S. Neutrino beams from muon storage rings: characteristics and physics potential. *Phys. Rev. D* **57**, 6989–6997 (1998).
- Alsharo'a, M. M. et al. Recent progress in neutrino factory and muon collider research within the Muon Collaboration. *Phys. Rev. Accel. Beams* **6**, 081001 (2003).
- Palmer, R. B. Muon colliders. *Rev. Accel. Sci. Tech.* **7**, 137–159 (2014).
- Boscolo, M. et al. Low emittance muon accelerator studies with production from positrons on target. *Phys. Rev. Accel. Beams* **21**, 061005 (2018).
- Neuffer, D. & Shiltsev, V. On the feasibility of a pulsed 14 TeV c.m.e. muon collider in the LHC tunnel. *J. Instrum.* **13**, T10003 (2018).
- Skrinsky, A. N. & Parkhomchuk, V. V. Cooling methods for beams of charged particles. *Sov. J. Part. Nucl.* **12**, 223–247 (1981).
- Neuffer, D. Principles and applications of muon cooling. *Part. Accel.* **14**, 75–90 (1983).
- Rogers, C. T. et al. Muon front end for the neutrino factory. *Phys. Rev. Accel. Beams* **16**, 040104 (2013).
- Stratakis, D. & Palmer, R. B. Rectilinear six-dimensional ionization cooling channel for a muon collider: a theoretical and numerical study. *Phys. Rev. Accel. Beams* **18**, 031003 (2015).
- Neuffer, D. et al. Final cooling for a high-energy high-luminosity lepton collider. *J. Instrum.* **12**, T07003 (2017).
- Lawrence, E. O. & Livingston, M. S. The production of high speed protons without the use of high voltages. *Phys. Rev.* **38**, 834 (1931).
- Lewis, G. N., Livingston, M. S. & Lawrence, E. O. The emission of alpha-particles from various targets bombarded by deuterons of high speed. *Phys. Rev.* **44**, 55–56 (1933).
- R. Wideröe. Das Betatron. *Z. Angew. Phys.* **5**, 187–200 (1953).
- Behnke, T. et al. *The International Linear Collider Technical Design Report – Volume 1: Executive Summary* (ILC, 2013).
- Burrows, P. N. et al. (eds) *The Compact Linear Collider (CLIC): 2018 Summary Report* (CERN, 2018).
- CEPC Study Group. *CEPC Conceptual Design Report: Volume 1 – Accelerator*. (IHEP, 2018).
- Abada, A. et al. FCC-ee: the lepton collider. *Eur. Phys. J. Spec. Top.* **228**, 261–623 (2019).
- Myers, S. The Large Hadron Collider 2008–2013. *Int. J. Mod. Phys. A* **28**, 1330035 (2013).
- Lee, S. Y. *Accelerator Physics* 3rd edn (World Scientific, 2012).
- Schröder, S. et al. First laser cooling of relativistic ions in a storage ring. *Phys. Rev. Lett.* **64**, 2901–2904 (1990).
- Möhl, D., Petrucci, G., Thorndahl, L., & van der Meer, S. Physics and technique of stochastic cooling. *Phys. Rep.* **58**, 73–119 (1980).
- Parkhomchuk, V. V. & Skrinsky, A. N. Electron cooling: 35 years of development. *Phys. Uspekhi* **43**, 433–452 (2000).
- Mühlbauer, M. et al. Frictional cooling: experimental results. *Hyperfine Interact.* **119**, 305–310 (1999).
- Abramowicz, H. et al. A muon collider scheme based on frictional cooling. *Nucl. Instrum. Methods Phys. Res. A* **546**, 356–375 (2005).
- Taqqi, D. Compression and extraction of stopped muons. *Phys. Rev. Lett.* **97**, 194801 (2006).
- Bae, S. et al. First muon acceleration using a radio frequency accelerator. *Phys. Rev. Accel. Beams* **21**, 050101 (2018).
- Mori, Y. et al. Neutron source with emittance recovery internal target. In *Proc. of the 23rd Particle Accelerator Conference (JACoW, 2009)*; <http://accelconf.web.cern.ch/AccelConf/PAC2009/papers/th4gac04.pdf>.
- Penn, G. & Wurtele, J. S. Beam envelope equations for cooling of muons in solenoid fields. *Phys. Rev. Lett.* **85**, 764–767 (2000).
- Holzer, E. B. Figure of merit for muon cooling – an algorithm for particle counting in coupled phase planes. *Nucl. Instrum. Methods Phys. Res. A* **532**, 270–274 (2004).
- Rogers, C. *Beam Dynamics in an Ionization Cooling Channel*. PhD thesis, Imperial College London (2008).
- Drielsma, F. *Measurement of the Increase in Phase Space Density of a Muon Beam through Ionization Cooling*. PhD thesis, Univ. Geneva (2018).
- Bogomilov, M. et al. The MICE muon beam on ISIS and the beam-line instrumentation of the Muon Ionization Cooling Experiment. *J. Instrum.* **7**, P05009 (2012).
- Adams, D. et al. Characterisation of the muon beams for the Muon Ionization Cooling Experiment. *Eur. Phys. J. C* **73**, 2582 (2013).
- Bogomilov, M. et al. Pion contamination in the MICE muon beam. *J. Instrum.* **11**, P03001 (2016).
- Booth, C. N. et al. The design and performance of an improved target for MICE. *J. Instrum.* **11**, P05006 (2016).
- Thomason, J. W. G. The ISIS Spallation Neutron and Muon Source – the first thirty-three years. *Nucl. Instrum. Methods Phys. Res. A* **917**, 61–67 (2019).
- Bayliss, V. et al. The liquid-hydrogen absorber for MICE. *J. Instrum.* **13**, T09008 (2018).
- Stratakis, D., Palmer, R. B. & Grote, D. P. Influence of space-charge fields on the cooling process of muon beams. *Phys. Rev. Accel. Beams* **18**, 044201 (2015).



40. Blackmore, V. et al. First particle-by-particle measurement of emittance in the Muon Ionization Cooling Experiment. *Eur. Phys. J. C* **79**, 257 (2019).
41. Bertoni, R. et al. The design and commissioning of the MICE upstream time-of-flight system. *Nucl. Instrum. Methods Phys. Res. A* **615**, 14–26 (2010).
42. Cremaldi, L. et al. A Cherenkov radiation detector with high density aerogels. *IEEE Trans. Nucl. Sci.* **56**, 1475–1478 (2009).
43. Ellis, M. et al. The design, construction and performance of the MICE scintillating fibre trackers. *Nucl. Instrum. Methods Phys. Res. A* **659**, 136–153 (2011).
44. Dobbs, A. et al. The reconstruction software for the MICE scintillating fibre trackers. *J. Instrum.* **11**, T12001 (2016).
45. Bertoni, R. et al. *The Construction of the MICE TOF2 Detector*. MICE Technical Note 254 (2010); <http://mice.iit.edu/micenotes/public/pdf/MICE0286/MICE0286.pdf>.
46. Adams, D. et al. Electron–Muon Ranger: performance in the MICE Muon Beam. *J. Instrum.* **10**, P12012 (2015).
47. Asfandiyarov, R. et al. The design and construction of the MICE Electron–Muon Ranger. *J. Instrum.* **11**, T10007 (2016).
48. Petroff, M. D. & Stapelbroek, M. G. Photon-counting solid-state photomultiplier. *IEEE Trans. Nucl. Sci.* **36**, 158–162 (1989).
49. Agostinelli, S. et al. GEANT4: a simulation toolkit. *Nucl. Instrum. Methods Phys. Res. A* **506**, 250–303 (2003).
50. Asfandiyarov, R. et al. MAUS: the MICE Analysis User Software. *J. Instrum.* **14**, T04005 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

#### MICE collaboration

M. Bogomilov<sup>1</sup>, R. Tsenov<sup>1</sup>, G. Vankova-Kirilova<sup>1</sup>, Y. P. Song<sup>2</sup>, J. Y. Tang<sup>2</sup>, Z. H. Li<sup>3</sup>, R. Bertoni<sup>4</sup>, M. Bonesini<sup>4</sup>, F. Chignoli<sup>4</sup>, R. Mazza<sup>4</sup>, V. Palladino<sup>5</sup>, A. de Bari<sup>6</sup>, D. Orestano<sup>7</sup>, L. Tortora<sup>7</sup>, Y. Kuno<sup>8</sup>, H. Sakamoto<sup>8,34</sup>, A. Sato<sup>8</sup>, S. Ishimoto<sup>9</sup>, M. Chung<sup>10</sup>, C. K. Sung<sup>10</sup>, F. Filthaut<sup>11,12</sup>, D. Jokovic<sup>13</sup>, D. Maletic<sup>13</sup>, M. Savic<sup>13</sup>, N. Jovancevic<sup>14</sup>, J. Nikolov<sup>14</sup>, M. Vretenar<sup>15</sup>, S. Ramberger<sup>15</sup>, R. Asfandiyarov<sup>16</sup>, A. Blondel<sup>16</sup>, F. Drielsma<sup>16</sup>, Y. Karadzhov<sup>16</sup>, S. Boyd<sup>17</sup>, J. R. Greis<sup>17</sup>, T. Lord<sup>17</sup>, C. Pidcott<sup>17,35</sup>, I. Taylor<sup>17,36</sup>, G. Charnley<sup>18</sup>, N. Collomb<sup>18</sup>, K. Dumbell<sup>18</sup>, A. Gallagher<sup>18</sup>, A. Grant<sup>18</sup>, S. Griffiths<sup>18</sup>, T. Hartnett<sup>18</sup>, B. Martlew<sup>18</sup>, A. Moss<sup>18</sup>, A. Muir<sup>18</sup>, I. Mullacrane<sup>18</sup>, A. Oates<sup>18</sup>, P. Owens<sup>18</sup>, G. Stokes<sup>18</sup>, P. Warburton<sup>18</sup>, C. White<sup>18</sup>, D. Adams<sup>19</sup>, V. Bayliss<sup>19</sup>, J. Boehm<sup>19</sup>, T. W. Bradshaw<sup>19</sup>, C. Brown<sup>19,20</sup>, M. Courthold<sup>19</sup>, J. Govans<sup>19</sup>, M. Hills<sup>19</sup>, J.-B. Lagrange<sup>19</sup>, C. Macwaters<sup>19</sup>, A. Nichols<sup>19</sup>, R. Preece<sup>19</sup>, S. Ricciardi<sup>19</sup>, C. Rogers<sup>19\*</sup>, T. Stanley<sup>19</sup>,

J. Tarrant<sup>19</sup>, M. Tucker<sup>19</sup>, S. Watson<sup>19,37</sup>, A. Wilson<sup>19</sup>, R. Bayes<sup>21,38</sup>, J. C. Nugent<sup>21</sup>, F. J. P. Soler<sup>21</sup>, G. T. Chatzitheodoridis<sup>21,22,23</sup>, A. J. Dick<sup>22,23</sup>, K. Ronald<sup>22,23</sup>, C. G. Whyte<sup>22,23</sup>, A. R. Young<sup>22,23</sup>, R. Gamet<sup>24</sup>, P. Cooke<sup>24</sup>, V. J. Blackmore<sup>25</sup>, D. Colling<sup>25</sup>, A. Dobbs<sup>25,39</sup>, P. Dornan<sup>25</sup>, P. Franchini<sup>25</sup>, C. Hunt<sup>25,40</sup>, P. B. Jurj<sup>25</sup>, A. Kurup<sup>25</sup>, K. Long<sup>25</sup>, J. Martyniak<sup>25</sup>, S. Middleton<sup>25,41</sup>, J. Pasternak<sup>25</sup>, M. A. Uchida<sup>25,42</sup>, J. H. Cobb<sup>26</sup>, C. N. Booth<sup>27</sup>, P. Hodgson<sup>27</sup>, J. Langlands<sup>27</sup>, E. Overton<sup>27,43</sup>, V. Pec<sup>27</sup>, P. J. Smith<sup>27</sup>, S. Wilbur<sup>27</sup>, M. Ellis<sup>20,44</sup>, R. B. S. Gardener<sup>20</sup>, P. Kyberd<sup>20</sup>, J. J. Nebrensky<sup>20</sup>, A. DeMello<sup>28</sup>, S. Gourlay<sup>28</sup>, A. Lambert<sup>28</sup>, D. Li<sup>28</sup>, T. Luo<sup>28</sup>, S. Prestemon<sup>28</sup>, S. Virostek<sup>28</sup>, M. Palmer<sup>29</sup>, H. Witte<sup>29</sup>, D. Adey<sup>30,45</sup>, A. D. Bross<sup>30</sup>, D. Bowring<sup>30</sup>, A. Liu<sup>30,46</sup>, D. Neuffer<sup>30</sup>, M. Popovic<sup>30</sup>, P. Rubinov<sup>30</sup>, B. Freemire<sup>31,46</sup>, P. Hanlet<sup>31,47</sup>, D. M. Kaplan<sup>31</sup>, T. A. Mohaya<sup>31,47</sup>, D. Rajaram<sup>31</sup>, P. Snopok<sup>31</sup>, Y. Torun<sup>31</sup>, L. M. Cremaldi<sup>32</sup>, D. A. Sanders<sup>32</sup>, D. J. Summers<sup>32</sup>, L. R. Coney<sup>33,48</sup>, G. G. Hanson<sup>33</sup> & C. Heidt<sup>33</sup>

<sup>1</sup>Department of Atomic Physics, St Kliment Ohridski University of Sofia, Sofia, Bulgaria.

<sup>2</sup>Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China. <sup>3</sup>Sichuan University, Chengdu, China. <sup>4</sup>Sezione INFN Milano Bicocca, Dipartimento di Fisica G. Occhialini, Milan, Italy. <sup>5</sup>Sezione INFN Napoli and Dipartimento di Fisica, Università Federico II, Complesso Universitario di Monte S. Angelo, Naples, Italy. <sup>6</sup>Sezione INFN Pavia and Dipartimento di Fisica, Pavia, Italy. <sup>7</sup>INFN Sezione di Roma Tre and Dipartimento di Matematica e Fisica, Università Roma Tre, Rome, Italy. <sup>8</sup>Osaka University, Graduate School of Science, Department of Physics, Toyonaka, Japan. <sup>9</sup>High Energy Accelerator Research Organization (KEK), Institute of Particle and Nuclear Studies, Tsukuba, Japan. <sup>10</sup>UNIST, Ulsan, South Korea. <sup>11</sup>Nikhef, Amsterdam, The Netherlands. <sup>12</sup>Radboud University, Nijmegen, The Netherlands. <sup>13</sup>Institute of Physics, University of Belgrade, Belgrade, Serbia. <sup>14</sup>Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia. <sup>15</sup>CERN, Geneva, Switzerland. <sup>16</sup>DPNC, Section de Physique, Université de Genève, Geneva, Switzerland. <sup>17</sup>Department of Physics, University of Warwick, Coventry, UK. <sup>18</sup>STFC Daresbury Laboratory, Daresbury, Cheshire, UK. <sup>19</sup>STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, UK. <sup>20</sup>Brunel University, Uxbridge, UK. <sup>21</sup>School of Physics and Astronomy, The University of Glasgow, Glasgow, UK. <sup>22</sup>SUPA and Department of Physics, University of Strathclyde, Glasgow, UK. <sup>23</sup>Cockcroft Institute, Daresbury Laboratory, Daresbury, UK. <sup>24</sup>Department of Physics, University of Liverpool, Liverpool, UK. <sup>25</sup>Department of Physics, Blackett Laboratory, Imperial College London, London, UK. <sup>26</sup>Department of Physics, University of Oxford, Oxford, UK. <sup>27</sup>Department of Physics and Astronomy, University of Sheffield, Sheffield, UK. <sup>28</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA. <sup>29</sup>Brookhaven National Laboratory, Upton, NY, USA. <sup>30</sup>Fermilab, Batavia, IL, USA. <sup>31</sup>Illinois Institute of Technology, Chicago, IL, USA. <sup>32</sup>University of Mississippi, Oxford, MS, USA. <sup>33</sup>University of California, Riverside, CA, USA. <sup>34</sup>Present address: RIKEN 2-1 Horosawa, Wako, Japan. <sup>35</sup>Present address: Department of Physics and Astronomy, University of Sheffield, Sheffield, UK. <sup>36</sup>Present address: Defence Science and Technology Laboratory, Salisbury, UK. <sup>37</sup>Present address: ATC, Royal Observatory Edinburgh, Edinburgh, UK. <sup>38</sup>Present address: Laurentian University, Sudbury, Ontario, Canada. <sup>39</sup>Present address: OPERA Simulation Software, Kidlington, UK. <sup>40</sup>Present address: CERN, Geneva, Switzerland. <sup>41</sup>Present address: School of Physics and Astronomy, University of Manchester, Manchester, UK. <sup>42</sup>Present address: Cavendish Laboratory, Cambridge, UK. <sup>43</sup>Present address: Arm, Sheffield, UK. <sup>44</sup>Present address: Westpac Group, Sydney, New South Wales, Australia. <sup>45</sup>Present address: Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China. <sup>46</sup>Present address: Euclid Techlabs, Bolingbrook, IL, USA. <sup>47</sup>Present address: Fermilab, Batavia, IL, USA. <sup>48</sup>Present address: European Spallation Source ERIC, Lund, Sweden. \*e-mail: [chris.rogers@stfc.ac.uk](mailto:chris.rogers@stfc.ac.uk)

### Characterization of beam brightness

In particle accelerators, the average beam brightness  $\bar{B}$  is defined as the beam current,  $I$ , passing through a transverse phase-space volume  $\mathcal{V}_4$  (ref. <sup>51</sup>)

$$\bar{B} = \frac{I}{\mathcal{V}_4} \quad (1)$$

The normalized r.m.s. emittance is often used as an indicator of the phase-space volume occupied by the beam and is given by<sup>29</sup>

$$\varepsilon_{\perp} = \frac{\sqrt[4]{|V|}}{m_{\mu}c} \quad (2)$$

where  $m_{\mu}$  is the muon mass and  $|V|$  is the determinant of the covariance matrix of the beam in the transverse phase space  $\mathbf{u} = (x, p_x, y, p_y)$ . The covariance matrix has elements  $v_{ij} = \langle u_i u_j \rangle - \langle u_i \rangle \langle u_j \rangle$ . The distribution of individual particle amplitudes also describes the volume of the beam in phase space.

The amplitude is defined by<sup>30</sup>

$$A_{\perp} = \varepsilon_{\perp} R^2(\mathbf{u}, \langle \mathbf{u} \rangle) \quad (3)$$

where  $R^2(\mathbf{u}, \mathbf{v})$  is the square of the distance between two points,  $\mathbf{u}$  and  $\mathbf{v}$ , in the phase space, normalized to the covariance matrix:

$$R^2(\mathbf{u}, \mathbf{v}) = (\mathbf{u} - \mathbf{v})^T V^{-1} (\mathbf{u} - \mathbf{v}) \quad (4)$$

The normalized r.m.s. emittance is proportional to the mean of the particle amplitude distribution. In the approximation that particles travel near the beam axis, and in the absence of cooling, the particle amplitudes and the normalized r.m.s. emittance are conserved quantities. If the beam is well described by a multivariate Gaussian distribution, then  $R^2$  is distributed according to a  $\chi^2$  distribution with four degrees of freedom, so the amplitudes are distributed according to

$$f(A_{\perp}) = \frac{A_{\perp}}{4\varepsilon_{\perp}^2} \exp\left(-\frac{A_{\perp}}{2\varepsilon_{\perp}}\right) \quad (5)$$

The rate of change of the normalized transverse emittance as the beam passes through an absorber is given approximately by<sup>8,29,31</sup>

$$\frac{d\varepsilon_{\perp}}{dz} \approx -\frac{\varepsilon_{\perp}}{\beta^2 E_{\mu}} \left| \frac{dE_{\mu}}{dz} \right| + \frac{\beta_{\perp} (13.6 \text{ MeV } c^{-1})^2}{2\beta^3 E_{\mu} m_{\mu} X_0} \quad (6)$$

where  $\beta c$  is the muon velocity,  $E_{\mu}$  is the muon energy,  $|dE_{\mu}/dz|$  is the mean energy loss per unit path length,  $X_0$  is the radiation length of the absorber and  $\beta_{\perp}$  is the transverse betatron function at the absorber<sup>29</sup>. The first term of this equation describes ‘cooling’ by ionization energy loss and the second term describes ‘heating’ by multiple Coulomb scattering. Equation (6) implies that there is an equilibrium emittance for which the emittance change is zero.

If the beam is well described by a multivariate Gaussian distribution both before and after cooling, then the downstream and upstream amplitude distributions  $f^d(A_{\perp})$  and  $f^u(A_{\perp})$  are related to the downstream and upstream emittances  $\varepsilon_{\perp}^d$  and  $\varepsilon_{\perp}^u$  by

$$\frac{f^d(A_{\perp})}{f^u(A_{\perp})} = \left( \frac{\varepsilon_{\perp}^u}{\varepsilon_{\perp}^d} \right)^2 \exp\left[ -\frac{A_{\perp}}{2} \left( \frac{1}{\varepsilon_{\perp}^d} - \frac{1}{\varepsilon_{\perp}^u} \right) \right] \quad (7)$$

In the experiment described in this paper, many particles do not travel near the beam axis. These particles experience effects from

optical aberrations, as well as geometrical effects such as scraping, in which high-amplitude particles outside the experiment’s aperture are removed from the beam. Scraping reduces the emittance of the ensemble and selectively removes those particles that scatter more than the rest of the ensemble. Optical aberrations and scraping introduce a bias in the change in r.m.s. emittance that occurs because of ionization cooling. In this work the distribution of amplitudes is studied. To expose the behaviour in the beam core, independently of aberrations affecting the beam tail,  $V$  and  $\varepsilon_{\perp}$  are recalculated for each amplitude bin, including particles that are in lower-amplitude bins and excluding particles that are in higher-amplitude bins. This results in a distribution that, in the core of the beam, is independent of scraping effects and spherical aberrations.

The change in phase-space density provides a direct measurement of the cooling effect. The  $k$ -nearest-neighbour algorithm provides a robust non-parametric estimator of the phase-space density of the muon ensemble<sup>32,34,52</sup>. The separation of pairs of muons is characterized by the normalized squared distance,  $R_{ij}^2(\mathbf{u}_i, \mathbf{u}_j)$ , between muons with positions  $\mathbf{u}_i$  and  $\mathbf{u}_j$ . A volume  $\mathcal{V}_{ik}$  is associated with each particle, which corresponds to the hypersphere that is centred on  $\mathbf{u}_i$  and intersects the  $k$ th nearest particle (that is, the particle that has the  $k$ th smallest  $R_{ij}$ ). The density,  $\rho_i$ , associated with the  $i$ th particle is estimated by

$$\rho_i(\mathbf{u}_i) = \frac{k}{n} \frac{1}{|V|^{1/2} \mathcal{V}_{ik}} = \frac{2k}{n\pi^2} \frac{1}{|V|^{1/2} R_{ik}^4} \quad (8)$$

where  $n$  is the number of particles in the ensemble. An optimal value for  $k$  is used,  $k = n^{4/(4+d)} = \sqrt{n}$ , with phase-space dimension  $d=4$  (ref. <sup>32</sup>).

### Data taking and reconstruction

Data were buffered in the front-end electronics and read out after each target actuation. Data storage was triggered by a coincidence of signals in the photomultiplier tubes (PMTs) serving a single scintillator slab in the upstream TOF station closest to the cooling channel (TOF1). The data recorded in response to a particular trigger are referred to as a ‘particle event’.

Each TOF station was composed of a number of scintillator slabs that were read out using a pair of PMTs, one mounted at each end of each slab. The reconstruction of the data began with the search for coincidences in the signals from the two PMTs serving any one slab in a TOF plane. Such coincidences are referred to as ‘slab hits’. ‘Space points’ were then formed from the intersection of slab hits in the  $x$  and  $y$  projections of each TOF station separately. The position and time at which a particle giving rise to the space point crossed the TOF station were then calculated using the slab position and the times measured in each of the PMTs. The relative timing of the two upstream TOF stations (TOF0 and TOF1) was calibrated relative to the measured time taken for electrons to pass between the two TOF detectors, on the assumption that they travelled at the speed of light.

Signals in the tracker readout were collected to reconstruct the helical trajectories (‘tracks’) of charged particles in the upstream and downstream trackers (TKU and TKD, respectively). Multiple Coulomb scattering introduced significant uncertainties in the reconstruction of the helical trajectory of tracks with a bending radius of less than 5 mm. For this class of track, the momentum was deduced by combining the tracker measurement with the measurements from nearby detectors. The track-fitting quality was characterized by the  $\chi^2$  per degree of freedom

$$\chi_{\text{df}}^2 = \frac{1}{n} \sum_i \frac{\delta x_i^2}{\sigma_i^2} \quad (9)$$

where  $\delta x_i$  is the distance between the fitted track and the measured signal in the  $i$ th tracker plane,  $\sigma_i$  is the resolution of the position measurement in the tracker planes and  $n$  is the number of planes that had

a signal used in the track reconstruction. Further details of the reconstruction and simulation may be found in ref. <sup>50</sup>.

### Beam selection

Measurements made in the instrumentation upstream of the absorber were used to select the input beam. The input beam (the upstream sample) was composed of events that satisfied the following criteria:

- Exactly one space point was found in TOF0 and TOF1 and exactly one track in TKU.

- The track in TKU had  $\chi^2_{df} < 8$  and was contained within the 150-mm fiducial radius over the full length of TKU.
- The track in TKU had a reconstructed momentum in the range 135–145 MeV  $c^{-1}$ , corresponding to the momentum acceptance of the cooling cell.
- The time-of-flight between TOF0 and TOF1 was consistent with that of a muon, given the momentum measured in TKU.
- The radius at which the track in TKU passed through the diffuser was smaller than the diffuser aperture.

The beam emerging from the cooling cell (the downstream sample) was characterized using the subset of the upstream sample that satisfied the following criteria:

- Exactly one track was found in TKD.
- The track in TKD had  $\chi^2_{df} < 8$  and was contained within the 150-mm fiducial radius of TKD over the full length of the tracker.

The same sample-selection criteria were used to select events from the simulation of the experiment, which included a reconstruction of the electronics signals expected for the simulated particles.

### Calculation of amplitudes

The amplitude distributions obtained from the upstream and downstream samples were corrected for the effects of the detector efficiency and resolution and for the migration of events between amplitude bins. The corrected number of events in a bin,  $N_i^{corr}$ , was calculated from the raw number of events,  $N_j^{raw}$ , using

$$N_i^{corr} = E_i \sum_j S_{ij} N_j^{raw} \quad (10)$$

where  $E_i$  is the efficiency correction factor and  $S_{ij}$  accounts for the detector resolution and event migration.  $E_i$  and  $S_{ij}$  were estimated from the simulation of the experiment. The uncorrected and corrected amplitude distributions for a particular configuration are shown in Extended Data Fig. 1. The correction is small relative to the ionization cooling effect, which is clear even in the uncorrected distributions.

It can be seen from equation (7) that in the limit of small amplitudes, and in the approximation that the beam is normally distributed in the phase-space variables, the ratio of the number of muons is equal to the ratio of the square of the emittances,

$$\lim_{A_{\perp} \rightarrow 0} \frac{f^d(A_{\perp})}{f^u(A_{\perp})} = \left( \frac{\varepsilon_{\perp}^u}{\varepsilon_{\perp}^d} \right)^2 \quad (11)$$

The ratio of  $f^d$  to  $f^u$  in the lowest-amplitude bin of Fig. 3, which is an approximation to this ratio, is listed in Extended Data Table 1.

### Data availability

The unprocessed and reconstructed data that support the findings of this study are publicly available on the GridPP computing Grid at <https://doi.org/10.17633/rd.brunel.3179644> (MICE unprocessed data) and <https://doi.org/10.17633/rd.brunel.5955850> (MICE reconstructed data). Source data for Figs. 3–5 and Extended Data Fig. 1 are provided with the paper.

Publications using MICE data must contain the following statement: “We gratefully acknowledge the MICE collaboration for allowing us access to their data. Third-party results are not endorsed by the MICE collaboration.”

### Code availability

The MAUS software<sup>50</sup> that was used to reconstruct and analyse the MICE data is available at <https://doi.org/10.17633/rd.brunel.8337542>. The analysis presented here used MAUS version 3.3.2.

51. Reiser, M. in *Theory and Design of Charged Particle Beams* 51–103 (John Wiley & Sons, 2008).
52. Mack, Y. P. & Rosenblatt, M. Multivariate k-nearest neighbor density estimates. *J. Multiv. Anal.* **9**, 1–15 (1979).

**Acknowledgements** The work described here was made possible by grants from the Science and Technology Facilities Council (UK), the Department of Energy and the National Science Foundation (USA), the Istituto Nazionale di Fisica Nucleare (Italy), the European Union under the European Union’s Framework Programme 7 (AIDA project, grant agreement number 262025; TIARA project, grant agreement number 261905; and EuCARD), the Japan Society for the Promotion of Science, the National Research Foundation of Korea (number NRF-2016R1A5A1013277), the Ministry of Education, Science and Technological Development of the Republic of Serbia, the Institute of High Energy Physics/Chinese Academy of Sciences fund for collaboration between the People’s Republic of China and the USA, and the Swiss National Science Foundation in the framework of the SCOPES programme. We gratefully acknowledge all sources of support. We are grateful for the support given to us by the staff of the STFC Rutherford Appleton and Daresbury laboratories. We acknowledge the use of Grid computing resources deployed and operated by GridPP in the UK, <http://www.gridpp.ac.uk/>.

**Author contributions** All authors contributed considerably to the design or construction of the apparatus or to the data taking or analysis described here.

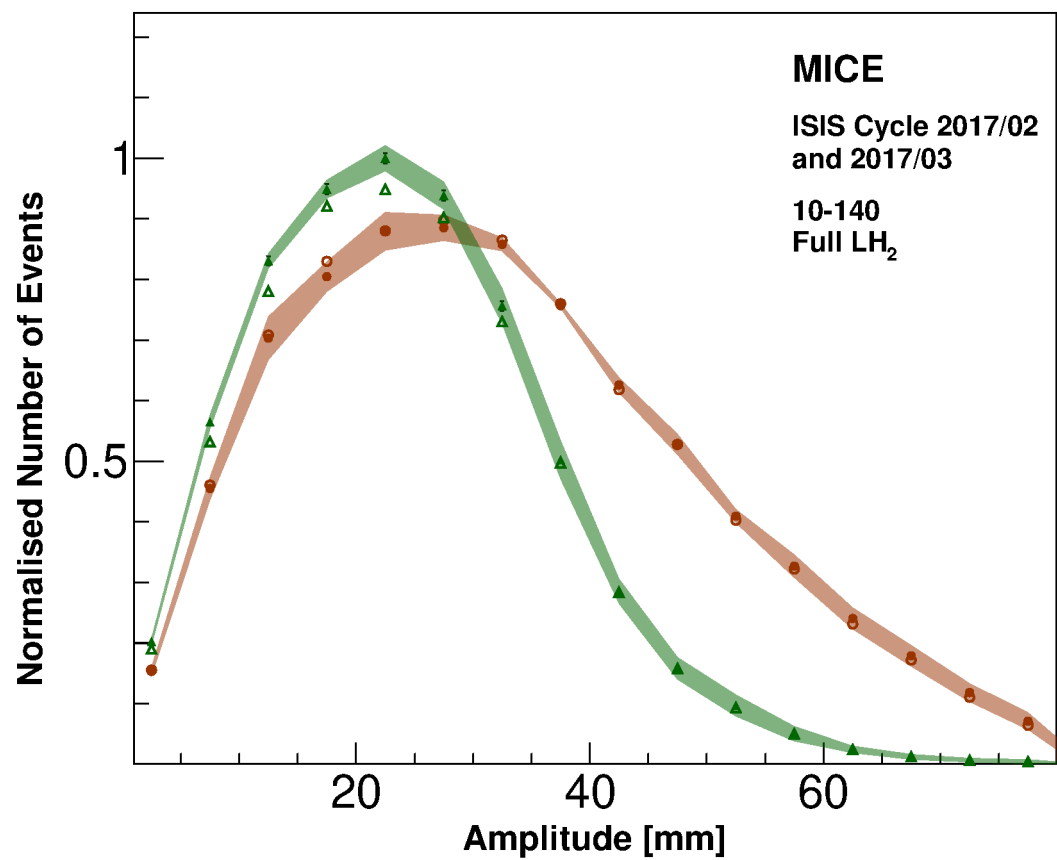
**Competing interests** The authors declare no competing interests.

### Additional information

**Correspondence and requests for materials** should be addressed to C.R.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





**Extended Data Fig. 1 | Corrected and uncorrected amplitude distributions for the 10–140 ‘LH2 full’ configuration.** The uncorrected data are shown by open points and the corrected data by filled points. Orange circles correspond to the upstream distribution and green triangles to the downstream distribution. Shading represents the estimated total standard error. Error bars show the statistical error and for most points are smaller than the markers.

**Extended Data Table 1 | Ratio of number of muons downstream to number of muons upstream having an amplitude of less than 5 mm**

	4-140	6-140	10-140
LH2 empty	0.98 ± 0.005 ± 0.05	1.01 ± 0.006 ± 0.05	1.02 ± 0.02 ± 0.05
LH2 full	1.12 ± 0.009 ± 0.05	1.16 ± 0.009 ± 0.05	1.32 ± 0.02 ± 0.07
None	0.99 ± 0.006 ± 0.05	1.00 ± 0.005 ± 0.05	1.06 ± 0.02 ± 0.06
LiH	1.07 ± 0.008 ± 0.05	1.14 ± 0.01 ± 0.05	1.23 ± 0.03 ± 0.07

Uncertainties denote standard error; statistical uncertainty is followed by the total uncertainty.

Extended Data Table 2 | Number of events in the samples shown in Fig. 3–5

	4-140		6-140		10-140	
	Upstream	Downstream	Upstream	Downstream	Upstream	Downstream
LH2 empty	163508	153813	158520	140981	123067	85082
LH2 full	71823	67640	117383	107329	82371	62660
None	91804	86877	172606	153809	54195	37436
LiH	87514	82682	98443	89875	43423	32715



# Coherent laser spectroscopy of highly charged ions using quantum logic

<https://doi.org/10.1038/s41586-020-1959-8>

Received: 16 July 2019

Accepted: 25 October 2019

Published online: 29 January 2020

P. Micke<sup>1,2,4\*</sup>, T. Leopold<sup>1,4</sup>, S. A. King<sup>1,4</sup>, E. Benkler<sup>1</sup>, L. J. Spieß<sup>1</sup>, L. Schmöger<sup>1,2</sup>, M. Schwarz<sup>1,2</sup>, J. R. Crespo López-Urrutia<sup>2</sup> & P. O. Schmidt<sup>1,3\*</sup>

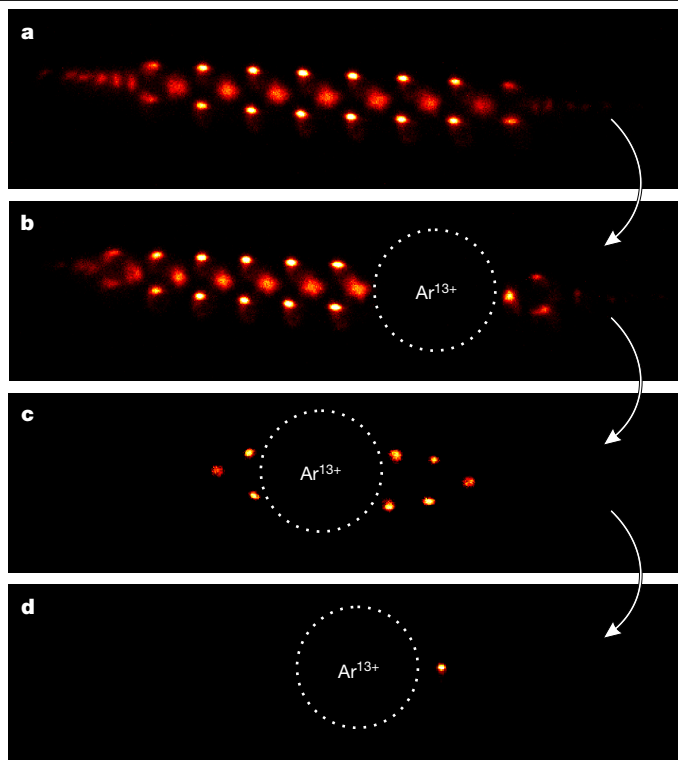
Precision spectroscopy of atomic systems<sup>1</sup> is an invaluable tool for the study of fundamental interactions and symmetries<sup>2</sup>. Recently, highly charged ions have been proposed to enable sensitive tests of physics beyond the standard model<sup>2–5</sup> and the realization of high-accuracy atomic clocks<sup>3,5</sup>, owing to their high sensitivity to fundamental physics and insensitivity to external perturbations, which result from the high binding energies of their outer electrons. However, the implementation of these ideas has been hindered by the low spectroscopic accuracies (of the order of parts per million) achieved so far<sup>6–8</sup>. Here we cool trapped, highly charged argon ions to the lowest temperature reported so far, and study them using coherent laser spectroscopy, achieving an increase in precision of eight orders of magnitude. We use quantum logic spectroscopy<sup>9,10</sup> to probe the forbidden optical transition in <sup>40</sup>Ar<sup>13+</sup> at a wavelength of 441 nanometres and measure its excited-state lifetime and *g*-factor. Our work unlocks the potential of highly charged ions as ubiquitous atomic systems for use in quantum information processing, as frequency standards and in highly sensitive tests of fundamental physics, such as searches for dark-matter candidates<sup>11</sup> or violations of fundamental symmetries<sup>2</sup>.

Like a microscope aimed at the quantum world, laser spectroscopy pursues ever higher resolving power. Every increase in resolution enables deeper insights into the subtle effects that all known fundamental interactions have on the atomic wavefunction. Advances in optical-frequency metrology have improved resolution drastically in the last three decades<sup>1</sup> and have made laser spectroscopy an extremely sensitive tool for studying open physics questions, such as the nature of dark matter, the strength of parity violation and a possible violation of Einstein's theory of relativity<sup>2</sup>. However, only a few atomic and ionic species are currently within the reach of cutting-edge optical-frequency metrology. Expanding this field of exploration to systems with high sensitivity to such effects is therefore crucial. Owing to the very high binding energies of their outer electrons, highly charged ions (HCIs) are promising candidates for such fundamental tests. The fractional contributions to the electronic transition energies from special relativity, quantum electrodynamics (QED) and the nucleus are several orders of magnitude larger than those in neutral atoms. This renders them ideal systems for benchmarking the most advanced theories and calculations, which has been repeatedly demonstrated via optical fluorescence spectroscopy in electron-beam ion traps (EBITs)<sup>6,7</sup>, X-ray spectroscopy in storage rings<sup>12</sup> and EBITs<sup>13–15</sup>, and ground-state *g*-factor studies in Penning traps<sup>16,17</sup>. The hyperfine splitting of the 1s state in heavy hydrogen-like ions can even shift into the optical range, providing laser-accessible transitions (see, for example, refs. <sup>18–20</sup>) with nuclear-size contributions of the order of several per cent of the total transition energy.

It was realized recently that non-gravitational coupling of dark matter to ordinary matter would affect atomic energy levels<sup>11</sup> and thus

become observable in optical-clock comparisons as an apparent drift or modulation of the fine-structure constant  $\alpha$ . HCIs offer narrow-linewidth optical transitions that are among the most sensitive to a possible variation of  $\alpha$  (ref. <sup>4</sup>). In addition, their inherent insensitivity to external electric fields<sup>3</sup> leads to considerably smaller systematic perturbations compared to neutral and singly charged atoms. This makes them potentially superior references for high-accuracy optical atomic clocks, with many proposed species reviewed in ref. <sup>5</sup>. However, so far no experiment has performed laser spectroscopy at the required level of precision. The major limitation was set by the high temperature of a few million kelvins at which HCIs are produced and typically stored. This induces Doppler broadenings with full-width-at-half-maximum (FWHM) linewidths of several tens of gigahertz and corresponding line-centre uncertainties of a few hundreds of megahertz in the best cases<sup>6–8</sup>. Because HCIs generally do not offer suitable transitions for direct laser cooling, sympathetic cooling of multiple HCIs by laser-cooled <sup>9</sup>Be<sup>+</sup> ions was implemented in a Penning trap at the Lawrence Livermore National Laboratory<sup>21</sup>, reaching an ion temperature of around 4 K. More recently, the Cryogenic Paul Trap Experiment<sup>22</sup> demonstrated reliable Coulomb crystallization of single <sup>40</sup>Ar<sup>13+</sup> ions in a crystal of many <sup>9</sup>Be<sup>+</sup> ions. Sympathetic Doppler cooling down to the 10-mK level and two-ion crystal preparation<sup>23,24</sup> paved the way for high-accuracy spectroscopy. Even so, spectroscopy of narrow transitions in single ions requires efficient state detection on a different, fast-cycling transition<sup>1</sup>, which is typically also used for laser cooling. If such a transition is not available, quantum logic spectroscopy (QLS) can be employed<sup>9,10</sup>. In QLS, the 'spectroscopy ion' (in

<sup>1</sup>Physikalisch-Technische Bundesanstalt, Braunschweig, Germany. <sup>2</sup>Max-Planck-Institut für Kernphysik, Heidelberg, Germany. <sup>3</sup>Institut für Quantenoptik, Leibniz Universität Hannover, Hannover, Germany. <sup>4</sup>These authors contributed equally: P. Micke, T. Leopold, S. A. King. \*e-mail: Peter.Micke@quantummetrology.de; Piet.Schmidt@quantummetrology.de



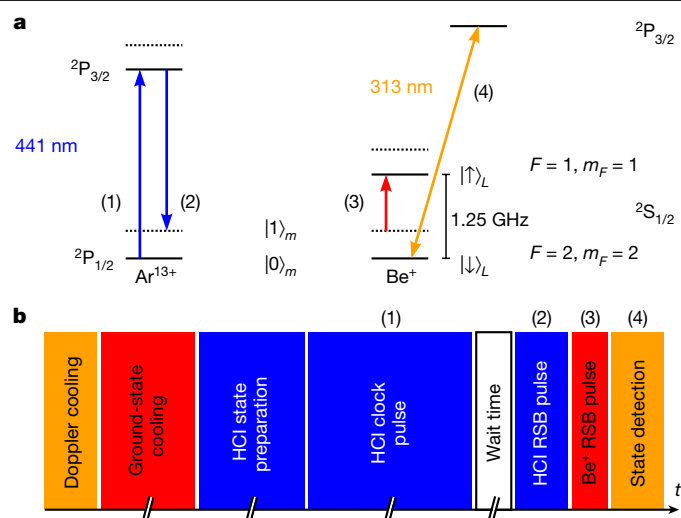
**Fig. 1 | Time sequence of HCI recapture and two-ion crystal preparation.** **a**, A laser-cooled Coulomb crystal of 50–100 fluorescing  ${}^9\text{Be}^+$  ions is confined in the Paul trap. **b**, A single  $\text{Ar}^{13+}$  ion is injected along the crystal axis, sympathetically cooled and finally co-crystallized with  ${}^9\text{Be}^+$ . It appears as a large dark void owing to the repulsion of the  ${}^9\text{Be}^+$  by the high charge state. **c**, Excess  ${}^9\text{Be}^+$  ions are removed by modulating the Paul trap radio-frequency potential in the absence of laser cooling, resulting in heating and ion losses. **d**, Finally, the  $\text{Ar}^{13+}$ – ${}^9\text{Be}^+$  two-ion crystal is prepared.

this case the HCI) is co-trapped with a so-called ‘logic ion’ ( ${}^9\text{Be}^+$ ) that provides sympathetic cooling and state preparation and is used for state detection. These functions are enabled by the Coulomb interaction between the two ions, allowing lasers to couple their internal electronic levels with their quantised joint motion in the trap. This technique and variations thereof have been successfully employed for optical atomic clocks based on  $\text{Al}^+$  ions<sup>25–27</sup>, for internal state detection and spectroscopy of molecular ions<sup>28,29</sup> and for spectroscopy of broad transitions in atomic ions<sup>30,31</sup>.

Here, we demonstrate QLS of an HCI—specifically, of the electric-dipole-forbidden transition between the  ${}^2\text{P}_{1/2}$  and  ${}^2\text{P}_{3/2}$  fine-structure levels of  ${}^{40}\text{Ar}^{13+}$  at a wavelength of 441 nm, the most accurately known transition in any HCI<sup>6</sup>. We achieve an FWHM well below 100 Hz, close to the natural linewidth of 17 Hz. Single line scans taken on a timescale of a few minutes determine the line centre with an uncertainty of less than 2 Hz. This corresponds to a fractional statistical uncertainty of  $3 \times 10^{-15}$  for a transition frequency of approximately 680 THz and compares favourably to previous measurements taken over hours or even days, which achieved relative uncertainties<sup>6–8</sup> of  $2 \times 10^{-7}$ . Quantum logic-assisted state preparation of the  ${}^{40}\text{Ar}^{13+}$  ion allows us to measure all six Zeeman components of the transition, which split up on a megahertz scale in a 160- $\mu\text{T}$  magnetic quantization field. This allows us to determine the  $g$ -factor of the  ${}^2\text{P}_{3/2}$  excited state with unprecedented accuracy. Furthermore, we demonstrate a quantum logic-assisted excited-state lifetime measurement.

### Preparation of a single HCI

A detailed description of the experimental setup is given in Methods (see also Extended Data Figs. 1, 2). In brief, argon HCIs are produced by

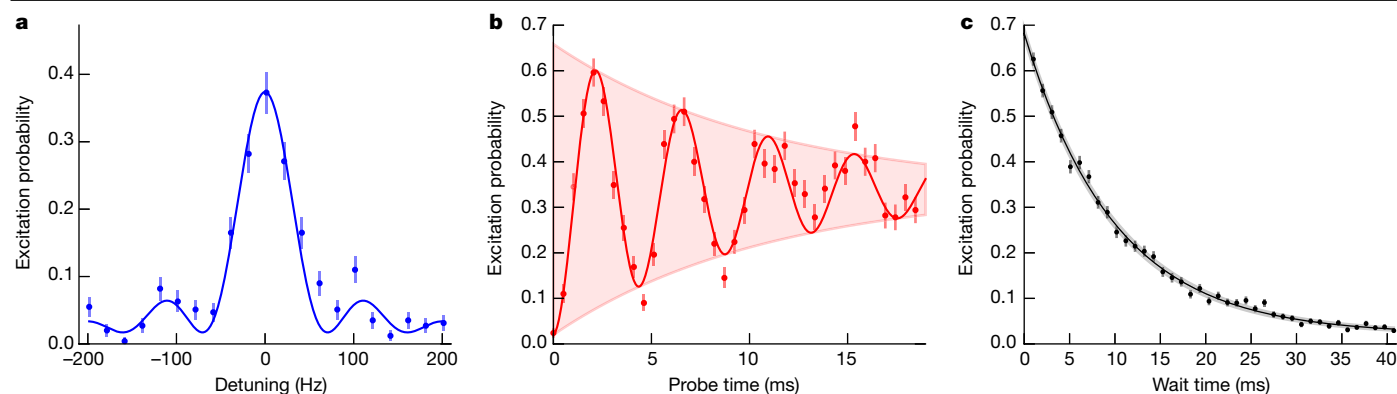


**Fig. 2 | Schematic illustration of the experimental cycle.** **a**, Level diagrams (not to scale) of boron-like  $\text{Ar}^{13+}$  and  ${}^9\text{Be}^+$ . The motional Fock state of the crystal is denoted as  $|n\rangle_m$ . Solid and dotted black lines indicate the corresponding ground and first excited motional states, respectively. **b**, Experimental sequence. After Doppler cooling on the  ${}^9\text{Be}^+ |\downarrow\rangle_L - {}^2\text{P}_{3/2}$  transition followed by ground-state cooling to  $|0\rangle_m$  by stimulated Raman transitions, the internal state of  $\text{Ar}^{13+}$  is prepared (see also Fig. 4c and Extended Data Fig. 3). A clock laser pulse addressing the carrier of the  $\text{Ar}^{13+}$  transition is then applied (1). After an optional wait time for lifetime measurements, a clock laser  $\pi$ -pulse on the  $\text{Ar}^{13+}$  red sideband (RSB) maps the  $\text{Ar}^{13+}$  electronic state onto the common motional state (2). A red-sideband  $\pi$ -pulse on the  ${}^9\text{Be}^+$  hyperfine transition  $|\downarrow\rangle_L - |\uparrow\rangle_L$  maps it onto the  ${}^9\text{Be}^+$  electronic state (3). Finally, this state is detected using the Doppler cooling laser (4).

an EBIT, PTB-EBIT<sup>32</sup>, and ejected from it in triggered bunches of ~200 ns duration with a mean kinetic energy of approximately 700 qV, where  $q$  is the ion charge. The HCIs are guided to the spectroscopy trap through an ion optical beamline. Based on their time of flight, we select the  ${}^{40}\text{Ar}^{13+}$  ions by rapidly switching a gate electrode. A pulsed gradient potential decelerates them electrostatically<sup>33</sup> to about 146 qV. Then, a single  ${}^{40}\text{Ar}^{13+}$  ion stochastically enters the cryogenic linear Paul trap<sup>34</sup>. The trap is globally biased to +138 V, thereby slowing the HCI down to 8 qV upon entry. After passing through the trapping region, the HCI is reflected back by an electrode at the end of the Paul trap. A mirror electrode in front of the trap is switched up to prevent the ion from escaping again, thereby capturing the HCI in an oscillatory axial motion. The repeated crossing through a pre-prepared laser-cooled  ${}^9\text{Be}^+$  Coulomb crystal within the trap dissipates the residual kinetic energy of the HCI. After sufficient sympathetic cooling, the  ${}^{40}\text{Ar}^{13+}$  ion joins the Coulomb crystal. Excess  ${}^9\text{Be}^+$  ions are removed until a two-ion crystal has been prepared (see Fig. 1). The entire preparation procedure of the two-ion crystal takes only a few minutes. The Paul trap is refrigerated to less than 5 K by a mechanically decoupled, closed-cycle cryostat to provide a vacuum below  $10^{-12}$  Pa (corresponding to a particle density of less than 20,000  $\text{cm}^{-3}$ ), thus suppressing charge-exchange collisions and achieving HCI storage times<sup>35</sup> of the order of 45 min.

### Ground-state cooling and quantum logic

The implementation of QLS requires control and preparation of the motional and internal states of both ions using coherent laser pulses on carrier and sideband transitions. After the two-ion crystal preparation, the strong Coulomb coupling between the two ions results in joint motional modes within the trap. Sympathetic cooling, state preparation and QLS are performed by repeating the experimental sequence shown in Fig. 2. First, Doppler cooling and optical pumping on the  ${}^9\text{Be}^+ {}^2\text{S}_{1/2} - {}^2\text{P}_{3/2}$  cycling transition (see Fig. 2a) are applied. The two axial normal



**Fig. 3 | Rabi spectroscopy and excited-state lifetime measurement. a**, Clock laser frequency scan across Zeeman component 1 (see Fig. 4c) of the  $^{40}\text{Ar}^{13+}$  fine-structure transition. The fixed probe time of 12 ms is longer than the excited-state lifetime of 9.6 ms. The line is fitted by a Rabi line shape (blue curve), reaching a Fourier-limited FWHM of about 65 Hz. **b**, On-resonance coherent excitation of this transition. The coherent state Rabi flopping signal (fitted by the red curve, which represents a damped sine with offset) exhibits a 2.2-ms  $\pi$ -time (in which the maximum transferable population is transferred to the excited state) and decays exponentially with the excited-state lifetime (red-

shaded envelope). The error bars in **a** and **b** represent the quantum projection noise of 255 measurements per data point. **c**, Excited-state lifetime measurement. Quantum logic sequences (see text) are carried out as a function of the wait time between carrier and red-sideband clock laser pulses. During the wait time, the excited state can decay spontaneously. From a three-parameter maximum-likelihood estimation, we obtain a lifetime of 9.97(27) ms, limited by the quantum projection noise of 1,100 measurements per data point (error bars). The black curve and grey-shaded area show the estimated exponential decay with the corresponding  $1\sigma$  uncertainty band.

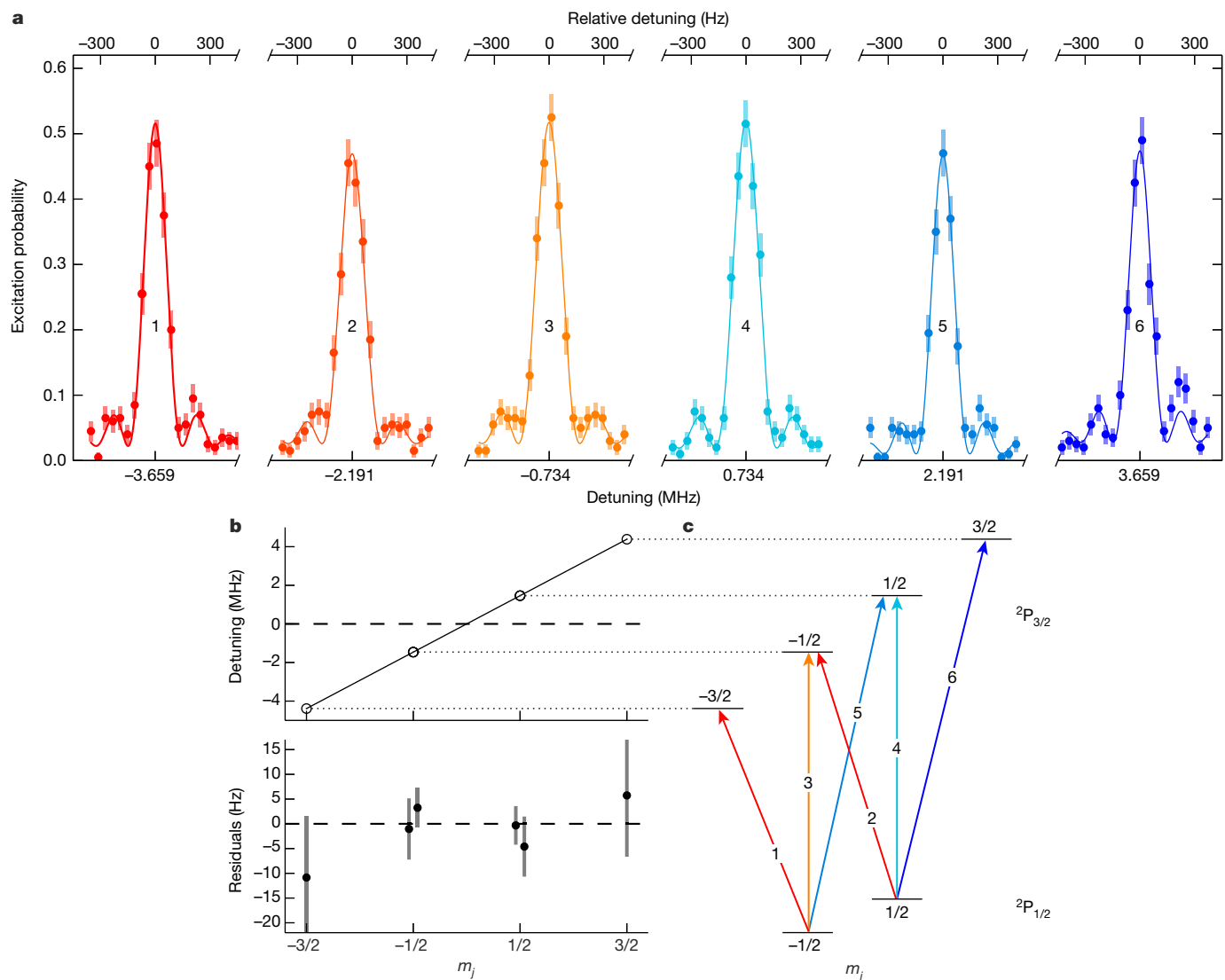
modes of the Coulomb crystal with secular frequencies of about  $\nu_{\text{IP}} = 1.37$  MHz (in phase) and  $\nu_{\text{OP}} = 1.86$  MHz (out of phase) are then cooled to the quantum mechanical ground state of motion with final average occupation numbers of  $\bar{n} = 0.05$  (in phase) and  $\bar{n} = 0.02$  (out of phase), corresponding to an effective temperature of less than 50  $\mu\text{K}$  for each mode. For this purpose, we use laser pulses that coherently couple the electronic degrees of freedom to the common motional modes (a technique referred to as resolved sideband cooling). To do this, stimulated Raman transitions are driven between the  $^9\text{Be}^+$  hyperfine qubit states  $|\downarrow\rangle_L |n\rangle_m \rightarrow |\uparrow\rangle_L |n-1\rangle_m$  (where  $|n\rangle_m$  denotes the motional quantum state of the in-phase or out-of-phase modes and  $|\uparrow\rangle_L$  refers to the upper qubit state of the logic ion  $^9\text{Be}^+$ ) using two laser beams with a wavelength of 313 nm. In our low magnetic field, the states are separated by a frequency of approximately 1.25 GHz. A repumping laser couples the  $^9\text{Be}^{+2}\text{S}_{1/2}$  and  $^2\text{P}_{1/2}$  levels (not shown in Fig. 2a) for electronic-state preparation and for depopulation of state  $|\uparrow\rangle_L$  (ref. <sup>36</sup>). The  $\text{Ar}^{13+}$  Zeeman ground state is then deterministically prepared with clock laser sideband pulses (see Extended Data Fig. 3). After full state preparation, QLS<sup>9</sup> is performed in four steps (see Fig. 2): (1) a clock laser pulse of tuneable length and power is applied, which couples the ground and excited states in  $\text{Ar}^{13+}$  coherently. (2) After a variable wait time for excited-state lifetime measurements, a clock laser red-sideband  $\pi$ -pulse maps the excitation from the electronic  $\text{Ar}^{13+}$  state onto the common axial out-of-phase mode and (3) another red-sideband  $\pi$ -pulse on the  $^9\text{Be}^+$  hyperfine qubit transition maps it onto the  $^9\text{Be}^+$  electronic state. (4) Finally, the qubit state of  $^9\text{Be}^+$  (dark,  $|\uparrow\rangle_L$ , or bright,  $|\downarrow\rangle_L$ ) is detected with a fidelity of up to 98% by counting the fluorescence photons that it scatters from the Doppler cooling laser within 200  $\mu\text{s}$ . A threshold value discriminates between the two states. The sequence is carried out multiple times (about 100) within a few seconds with a fixed set of parameters to average the quantum projection noise and evaluate a mean excitation probability. To resolve linewidths approaching the natural linewidth of 17 Hz, a narrow-linewidth clock laser is required. Our home-built laser system is composed of a commercial extended-cavity diode laser (ECDL) at 882 nm, which is prestabilized with a high locking bandwidth of 4 MHz to a passive external reference cavity using the ECDL pump current and grating piezo as actuators for the feedback. Thereby, we suppress laser high-frequency noise and obtain an instantaneous linewidth of about 2 kHz, which is limited by the relatively low cavity finesse (of about 1,000) and lack of vibration-insensitive design.

To suppress the residual noise, the laser is then further stabilized by phase-locking it to an ultrastable laser operating at a wavelength of 1.5  $\mu\text{m}$ . The latter is itself stabilized to a cryogenic cavity made from crystalline silicon (referred to as Si2)<sup>37</sup>. This achieves a fractional frequency instability at the thermal noise limit of the cavity of  $4 \times 10^{-17}$  at averaging times of 1–50 s. Using a femtosecond optical frequency comb as a transfer oscillator, we generate a virtual beatnote between the two lasers<sup>38</sup>. By demodulating it, we register their relative frequency and phase fluctuations, which are dominated by the substantially higher noise level of the 882-nm prestabilization cavity. The demodulated beatnote is used to generate a feedback signal for phase-locking the two lasers, which is applied to an acousto-optic modulator between the ECDL and the prestabilization cavity. The considerably lower bandwidth of the second locking stage ensures that the two loops do not compete with one another, but drifts and noise on the prestabilization cavity of up to kilohertz level are suppressed at the ECDL output, from which the spectroscopy light is derived. This suppresses the residual noise of the 882-nm clock laser, narrows its linewidth and reduces the daily drift to a level of about 10 Hz, which is dictated by Si2. The laser is frequency-doubled to 441 nm in an external enhancement cavity containing a periodically poled potassium titanyl phosphate crystal. Active power stabilization on a pulse-by-pulse basis at the ion trap is implemented for the clock, Doppler cooling and Raman lasers to achieve stable system parameters such as Rabi frequencies and a.c. Stark shifts.

### Coherent laser spectroscopy of $^{40}\text{Ar}^{13+}$

By applying this technique, we carried out the first coherent laser spectroscopy of an HCl. Figure 3a shows the excitation profile of the  $m_{1/2} = -1/2$  to  $m_{3/2} = -3/2$  Zeeman component of the  $^{40}\text{Ar}^{13+} 2\text{P}_{1/2} \rightarrow 2\text{P}_{3/2}$  fine-structure transition. The blue curve shows a fit to the line by a Rabi line shape, as expected for the top-hat laser pulse of 12-ms duration. This pulse length results in a Fourier-limited linewidth of 65 Hz FWHM. We do not observe any additional line broadenings on this first-order Zeeman sensitive transition at this level, which confirms our previous measurements<sup>34</sup> of a magnetic-field stability better than 1 nT achieved via active stabilization of the field in the vicinity of the vacuum chamber. Additionally, alternating external magnetic fields are shielded by the highly conductive cryogenic thermal shields made of high-purity copper with a low-pass corner frequency of  $\leq 0.3$  Hz and suppression



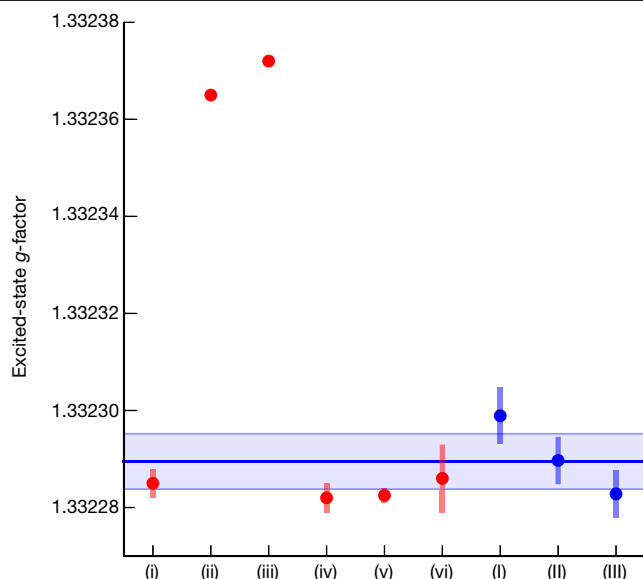


**Fig. 4 | Zeeman structure of the  $^{40}\text{Ar}^{13+} 2P_{1/2} - 2P_{3/2}$  fine-structure transition.** **a**, Excitation probability as a function of the clock laser detuning from the degenerate line centre, showing the six individual Zeeman components. Comparable laser-ion couplings ( $\pi$ -times of 5–6 ms) were chosen for each component, corresponding to Fourier-limited linewidths of around 150 Hz. The solid curves are Rabi line-shape fitting functions used to determine the centre frequencies. Error bars represent the quantum projection noise of 200 repetitions. The varying excitation probabilities of the six components are caused by slightly different state preparation efficiencies. **b**, Reconstructed

Zeeman shifts of the  $2P_{3/2}$  substates (upper panel) and their residuals (lower panel) with respect to a linear fit (solid black line). Note the different vertical scales. For each of the substates  $m_j = \pm 1/2$ , two data points are obtained from transitions 2, 3 and 4, 5, respectively. A magnetic-field instability of about 0.5 nT contributes to the standard uncertainties of the line centres, which become larger for the outer components. The slope of the linear fit is proportional to the ratio of the  $g$ -factors of the excited and ground states. **c**, Level diagram of the  $2P_{1/2}$  and  $2P_{3/2}$  Zeeman substates and corresponding Zeeman components of the fine-structure transition.

of 30–40 dB in the frequency range 60 Hz–1 kHz (ref. <sup>34</sup>). The maximum fringe contrast of about 0.4 at this probe duration was mostly limited by the excited-state lifetime, with contributions from the ~90% fidelity of the sideband operations on the two ions, as well as from imperfect state preparation and detection. Frequency scans with longer probe times can in principle resolve the natural linewidth, albeit at a reduced excitation probability. Figure 3b shows the on-resonance excitation probability as a function of the probe time for a higher intensity of the clock laser. Under continuous illumination, Rabi flopping between the two electronic states is observed (fitted by the red curve). The coherence decays with the known excited-state lifetime<sup>39</sup> of  $9.573^{(+4)}_{(-4)} \text{stat}^{(+12)}_{(-5)} \text{syst}$  ms (where the statistical and systematic standard deviations are given in parentheses), indicated by the red-shaded exponential envelope of the fit. This measurement confirms coherence beyond this timescale for both the clock laser and the magnetic field.

We also performed a direct measurement of the excited-state lifetime. For this, a carrier  $\pi$ -pulse (step (1) in Fig. 2) with maximum laser intensity was applied, which populated the  $^{40}\text{Ar}^{13+}$  excited state in about 16  $\mu\text{s}$ . After a variable wait time, the full transfer sequence was performed, and the remaining  $^{40}\text{Ar}^{13+}$  excited-state fraction was mapped onto the  $^9\text{Be}^+$  qubit state (see also Methods). During the wait time, a series of ground-state-cooling pulses on both axial motional modes was applied every millisecond to keep the two-ion crystal in the motional ground state in the presence of anomalous heating of 12 and 29 phonons per second for the out-of-phase and in-phase modes, respectively. By incrementing the wait time in 1-ms steps, an axial mode temperature independent of the wait time was ensured. The observed exponential spontaneous decay of the excited state is shown in Fig. 3c and results in a lifetime of 9.97(27) ms. This is about 1.5 standard deviations longer than the more accurate experimental result of



**Fig. 5 | Comparison of calculated (red) and measured (blue) excited-state  $g$ -factors.** Shown are results from Glazov et al.<sup>42</sup> (i), Verdebout et al.<sup>43</sup> (ii), Marques et al.<sup>44</sup> (iii), Shchepetnov et al.<sup>46</sup> (iv), Agababaev et al.<sup>45</sup> (v) and Maison et al.<sup>47</sup> (vi). The error bars of (iii) and (v) are smaller than the data points. No uncertainty is provided for data point (ii). (I), (II) and (III) represent  $g$ -factors evaluated from the three datasets produced in this work with their standard uncertainty. The solid blue line displays the weighted average with the  $1\sigma$  uncertainty band, with the largest contribution coming from the systematic uncertainty. See Methods for details.

$9.573(^{+4}_{-4})_{\text{stat}}(^{+12}_{-5})_{\text{syst}}$  ms obtained from an in-EBIT measurement<sup>39</sup> and than advanced calculations of 9.538(2) ms (ref.<sup>40</sup>) and 9.5354(20) ms (ref.<sup>41</sup>). Our result is consistent with the previous measurement and calculations within the uncertainty. Further details are provided in Methods.

### Measurement of the excited-state $g$ -factor

A magnetic field of about 160  $\mu\text{T}$  is applied at the location of the ions to define a quantization axis and to deliberately split the Zeeman substates of  $^9\text{Be}^+$  and  $^{40}\text{Ar}^{13+}$  on the megahertz scale. With quantum logic-assisted HCI state preparation (see Fig. 2b and Extended Data Fig. 3), all six Zeeman components of the  $^{40}\text{Ar}^{13+} 2\text{P}_{1/2} \rightarrow 2\text{P}_{3/2}$  transition can be coherently excited, as shown in Fig. 4a. The bottom horizontal axes represent the clock laser detuning from the degenerate line centre, and the top horizontal axes represent the relative detunings from the centres of the individual Zeeman components. We can reconstruct the Zeeman shifts of the  $2\text{P}_{3/2}$  substates from these data (see Fig. 4b, c and Methods) and derive the ratio of the  $g$ -factors of the excited and ground states from the measured frequencies. Within our current experimental precision of a few hertz over a splitting of several megahertz, we do not observe any quadratic contribution, which for instance could arise from coupling of electric field gradients to the electric quadrupole moment of the  $2\text{P}_{3/2}$  state, or from a quadratic Zeeman shift. A quadratic term added to the fitting function of Fig. 4b is consistent with zero. Recently, the ground-state  $g$ -factor  $g_{1/2}$  of  $^{40}\text{Ar}^{13+}$  was measured in the Penning trap experiment ALPHATRAP with an accuracy of parts per billion using the continuous Stern–Gerlach method<sup>17</sup>. Using this value, we obtain a weighted average of  $g_{3/2} = 1.3322895(13)_{\text{stat}}(56)_{\text{syst}}$  from three individual measurements (see Fig. 5). This is an improvement of more than two orders of magnitude over previous in-EBIT measurements<sup>7</sup>, revealing the contributions that arise from special relativity, interelectronic interactions and QED to an HCI excited-state  $g$ -factor. It also settles a discrepancy between previous theoretical values<sup>42–47</sup>, confirming the configuration-interaction calculations of refs.<sup>42,45,46</sup> and very recent coupled-cluster calculations<sup>47</sup>.

### Conclusions

We have cooled HCIs to the ground state of motion in a linear Paul trap, making them the coldest HCIs prepared in a laboratory so far. This enabled us to perform coherent, optical-clock-like laser spectroscopy of an electric-dipole-forbidden optical transition in an HCI using quantum logic, at a level of precision that is eight orders of magnitude higher than the previous state of the art. This proves the feasibility of hertz-level optical spectroscopy of HCIs and opens up this large class of atomic systems to the tools of cutting-edge frequency metrology and quantum information processing.

The determination of the absolute frequency of the  $^{40}\text{Ar}^{13+}$  fine-structure transition with a fractional uncertainty of  $3 \times 10^{-15}$  and even higher levels of precision requires further evaluation of systematic shifts, such as the small time dilation shift from the residual motion of the ion<sup>48</sup> or the electric quadrupole shift<sup>49</sup>, which is typically suppressed in HCIs. By restricting measurements to the points of maximum frequency sensitivity of each line, frequency information can be obtained faster than when scanning the full line profiles, as demonstrated here, further reducing the statistical uncertainty at a given averaging time<sup>50</sup>. At the same time, averaging over the Zeeman components on second—rather than minute—timescales will suppress systematic uncertainties arising from drifting magnetic fields<sup>51</sup>.

The presented techniques are not limited to our proof-of-principle HCI,  $^{40}\text{Ar}^{13+}$ , but can be applied more generally to forbidden transitions in other HCIs. Several of the candidate species have properties that are even better suited for optical-clock experiments, including much longer excited-state lifetimes and suppressed systematic shifts. Certain HCIs are particularly sensitive to physics beyond the standard model, such as possible variations of the fine-structure constant<sup>4</sup>, or to effects arising from fundamental interactions. Particularly, HCIs allow the systematic study of relativistic effects in bound electronic systems and of bound-state QED along isoelectronic sequences at ultrahigh precision<sup>5</sup>.

Furthermore, the techniques that we have demonstrated here are not limited to the optical domain. Our work also unlocks the new frontiers of the vacuum ultraviolet and X-ray regimes for ultrahigh precision spectroscopy—regions of the electromagnetic spectrum that are incompatible with neutral and singly charged atoms owing to unavoidable photoionization. This will enable novel high-accuracy atomic clocks based on HCIs and unrivalled tests of fundamental physics.

We note that during the revision of the manuscript, a complementary work demonstrating incoherent laser spectroscopy of  $^{40}\text{Ar}^{13+}$  in a Penning trap was published<sup>52</sup>.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1959-8>.

- Ludlow, A. D., Boyd, M. M., Ye, J., Peik, E. & Schmidt, P. O. Optical atomic clocks. *Rev. Mod. Phys.* **87**, 637–701 (2015).
- Safronova, M. S. et al. Search for new physics with atoms and molecules. *Rev. Mod. Phys.* **90**, 025008 (2018).
- Schiller, S. Hydrogenlike highly charged ions for tests of the time independence of fundamental constants. *Phys. Rev. Lett.* **98**, 180801 (2007).
- Berengut, J., Dzuba, V. & Flambaum, V. Enhanced laboratory sensitivity to variation of the fine-structure constant using highly charged ions. *Phys. Rev. Lett.* **105**, 120801 (2010).
- Kozlov, M. G., Safronova, M. S., Crespo López-Urrutia, J. R. & Schmidt, P. O. Highly charged ions: optical clocks and applications in fundamental physics. *Rev. Mod. Phys.* **90**, 045005 (2018).
- Draganić, I. et al. High precision wavelength measurements of QED-sensitive forbidden transitions in highly charged argon ions. *Phys. Rev. Lett.* **91**, 183001 (2003).
- Soria Orts, R. et al. Zeeman splitting and  $g$  factor of the  $1s^2 2s^2 2p^2 \text{P}_{3/2}$  and  $^2\text{P}_{3/2}$  levels in  $\text{Ar}^{13+}$ . *Phys. Rev. A* **76**, 052501 (2007).

8. Mäckel, V., Klawitter, R., Brenner, G., Crespo López-Urrutia, J. R. & Ullrich, J. Laser spectroscopy on forbidden transitions in trapped highly charged  $\text{Ar}^{13+}$  ions. *Phys. Rev. Lett.* **107**, 143002 (2011).
9. Schmidt, P. O. et al. Spectroscopy using quantum logic. *Science* **309**, 749–752 (2005).
10. Wineland, D. J., Bergquist, J. C., Bollinger, J. J., Drullinger, R. E. & Itano, W. M. Quantum computers and atomic clocks. In *Proc. 6th Symp. on Frequency Standards and Metrology* (ed. Gill, P.) 361–368 (World Scientific, 2002).
11. Derevianko, A. & Pospelov, M. Hunting for topological dark matter with atomic clocks. *Nat. Phys.* **10**, 933–936 (2014).
12. Gumberidze, A. et al. Quantum electrodynamics in strong electric fields: the ground-state Lamb shift in hydrogenlike uranium. *Phys. Rev. Lett.* **94**, 223001 (2005).
13. Beiersdorfer, P., Osterheld, A. L., Scofield, J. H., Crespo López-Urrutia, J. R. & Widmann, K. Measurement of QED and hyperfine splitting in the  $2s_{1/2}$ – $2p_{3/2}$  X-ray transition in Li-like  $^{209}\text{Bi}^{80+}$ . *Phys. Rev. Lett.* **80**, 3022–3025 (1998).
14. Beiersdorfer, P., Chen, H., Thorn, D. B. & Träbert, E. Measurement of the two-loop Lamb shift in lithiumlike  $\text{U}^{89+}$ . *Phys. Rev. Lett.* **95**, 233003 (2005).
15. Beiersdorfer, P. et al. Hyperfine splitting of the  $2s_{1/2}$  and  $2p_{1/2}$  levels in Li- and Be-like ions of  $^{141}\text{Pr}$ . *Phys. Rev. Lett.* **112**, 233003 (2014).
16. Sturm, S. et al.  $g$  factor of hydrogenlike  $^{28}\text{Si}^{13+}$ . *Phys. Rev. Lett.* **107**, 023002 (2011).
17. Arapoglou, I. et al.  $g$ -factor of boronlike argon  $^{40}\text{Ar}^{13+}$ . *Phys. Rev. Lett.* **122**, 253001 (2019).
18. Crespo López-Urrutia, J. R., Beiersdorfer, P., Savin, D. W. & Widmann, K. Direct observation of the spontaneous emission of the hyperfine transition  $F = 4$  to  $F = 3$  in ground state hydrogenlike  $^{165}\text{Ho}^{66+}$  in an electron beam ion trap. *Phys. Rev. Lett.* **77**, 826–829 (1996).
19. Seelig, P. et al. Ground state hyperfine splitting of hydrogenlike  $^{207}\text{Pb}^{81+}$  by laser excitation of a bunched ion beam in the GSI experimental storage ring. *Phys. Rev. Lett.* **81**, 4824–4827 (1998).
20. Ullmann, J. et al. High precision hyperfine measurements in Bismuth challenge bound-state strong-field QED. *Nat. Commun.* **8**, 15484 (2017).
21. Gruber, L. et al. Evidence for highly charged ion Coulomb crystallization in multicomponent strongly coupled plasmas. *Phys. Rev. Lett.* **86**, 636–639 (2001).
22. Schwarz, M. et al. Cryogenic linear Paul trap for cold highly charged ion experiments. *Rev. Sci. Instrum.* **83**, 083115–083115–10 (2012).
23. Schmöger, L. et al. Coulomb crystallization of highly charged ions. *Science* **347**, 1233–1236 (2015).
24. Schmöger, L. *Kalte hochgeladene Ionen für Frequenzmetrologie*. PhD thesis, Univ. of Heidelberg (2017).
25. Rosenband, T. et al. Frequency ratio of  $\text{Al}^+$  and  $\text{Hg}^+$  single-ion optical clocks; metrology at the  $17^{\text{th}}$  decimal place. *Science* **319**, 1808–1812 (2008).
26. Chou, C. W., Hume, D. B., Koelemeij, J. C. J., Wineland, D. J. & Rosenband, T. Frequency comparison of two high-accuracy  $\text{Al}^+$  optical clocks. *Phys. Rev. Lett.* **104**, 070802 (2010).
27. Brewer, S. M. et al.  $^{27}\text{Al}^+$  quantum-logic clock with a systematic uncertainty below  $10^{-18}$ . *Phys. Rev. Lett.* **123**, 033201 (2019).
28. Wolf, F. et al. Non-destructive state detection for quantum logic spectroscopy of molecular ions. *Nature* **530**, 457–460 (2016).
29. Chou, C. et al. Preparation and coherent manipulation of pure quantum states of a single molecular ion. *Nature* **545**, 203–207 (2017).
30. Hempel, C. et al. Entanglement-enhanced detection of single-photon scattering events. *Nat. Photon.* **7**, 630–633 (2013).
31. Wan, Y. et al. Precision spectroscopy by photon-recoil signal amplification. *Nat. Commun.* **5**, 3096 (2014).
32. Micke, P. et al. The Heidelberg compact electron beam ion traps. *Rev. Sci. Instrum.* **89**, 063109 (2018).
33. Schmöger, L. et al. Deceleration, precooling, and multi-pass stopping of highly charged ions in  $\text{Be}^+$  Coulomb crystals. *Rev. Sci. Instrum.* **86**, 103111 (2015).
34. Leopold, T. et al. A cryogenic radio-frequency ion trap for quantum logic spectroscopy of highly charged ions. *Rev. Sci. Instrum.* **90**, 073201 (2019).
35. Micke, P. et al. Closed-cycle, low-vibration 4 K cryostat for ion traps and other applications. *Rev. Sci. Instrum.* **90**, 065104 (2019).
36. King, S. A., Leopold, T., Thekkeppatt, P. & Schmidt, P. O. A self-injection locked DBR laser for laser cooling of beryllium ions. *Appl. Phys. B* **124**, 214 (2018).
37. Matei, D. G. et al. 1.5  $\mu\text{m}$  lasers with sub-10 mHz linewidth. *Phys. Rev. Lett.* **118**, 263202 (2017).
38. Stenger, J., Schnatz, H., Tamm, C. & Telle, H. Ultraprecise measurement of optical frequency ratios. *Phys. Rev. Lett.* **88**, 073601 (2002).
39. Lapiere, A. et al. Lifetime measurement of the  $\text{Ar XIV } 1s^2 2s^2 2p^2 P^{\circ}_{3/2}$  metastable level at the Heidelberg electron-beam ion trap. *Phys. Rev. A* **73**, 052507 (2006).
40. Tupitsyn, I. I. et al. Magnetic-dipole transition probabilities in B-like and Be-like ions. *Phys. Rev. A* **72**, 062503 (2005).
41. Bilal, M., Volotka, A. V., Beerwerth, R. & Fritzsche, S. Line strengths of QED-sensitive forbidden transitions in B-, Al-, F- and Cl-like ions. *Phys. Rev. A* **97**, 052506 (2018).
42. Glazov, D. A. et al.  $g$  factor of boron-like ions: ground and excited states. *Phys. Scr.* **T156**, 014014 (2013).
43. Verdebout, S. et al. Hyperfine structures and Landé  $g$ -factors for  $n = 2$  states in beryllium-, boron-, carbon-, and nitrogen-like ions from relativistic configuration interaction calculations. *At. Data Nucl. Data Tables* **100**, 1111–1155 (2014).
44. Marques, J. P., Indelicato, P., Parente, F., Sampaio, J. M. & Santos, J. P. Ground-state Landé  $g$  factors for selected ions along the boron isoelectronic sequence. *Phys. Rev. A* **94**, 042504 (2016).
45. Agababae, V. A. et al.  $g$  factor of the  $[(1s)^2(2s)^22p]^2P_{3/2}$  state of middle-Z boronlike ions. *X-ray Spectrom.* **49**, 143–148 (2019).
46. Shchepetnov, A. A. et al. Nuclear recoil correction to the  $g$  factor of boron-like argon. *J. Phys. Conf. Ser.* **583**, 012001 (2015).
47. Maison, D. E., Skripnikov, L. V. & Glazov, D. A. Many-body study of the  $g$  factor in boronlike argon. *Phys. Rev. A* **99**, 042506 (2019).
48. Berkeland, D. J., Miller, J. D., Bergquist, J. C., Itano, W. M. & Wineland, D. J. Minimization of ion micromotion in a Paul trap. *J. Appl. Phys.* **83**, 5025–5033 (1998).
49. Itano, W. M. External-field shifts of the  $^{199}\text{Hg}^+$  optical frequency standard. *J. Res. Natl. Inst. Stand. Technol.* **105**, 829–837 (2000).
50. Madej, A. A., Dubé, P., Zhou, Z., Bernard, J. E. & Gertsolf, M.  $^{88}\text{Sr}^+$  445-THz single-ion reference at the  $10^{-17}$  level via control and cancellation of systematic uncertainties and its measurement against the SI second. *Phys. Rev. Lett.* **109**, 203002 (2012).
51. Barwood, G. P., Huang, G., King, S. A., Klein, H. A. & Gill, P. Frequency noise processes in a strontium ion optical clock. *J. Phys. At. Mol. Opt. Phys.* **48**, 035401 (2015).
52. Egl, A. et al. Application of the continuous Stern–Gerlach effect for laser spectroscopy of the  $^{40}\text{Ar}^{13+}$  fine structure in a Penning trap. *Phys. Rev. Lett.* **123**, 123001 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020



### HCI production, transfer and recapture

We show a top view of the laboratory setup in Extended Data Fig. 1 and a simplified schematic of the potential landscape in Extended Data Fig. 2a. HCIs are produced by electron impact ionization and stored by PTB-EBIT, a Heidelberg-type compact EBIT<sup>32</sup>. After extraction of ions in bunches, a beamline with multiple electrostatic elements is used to guide the ions towards the Paul trap and to manipulate their kinetic-energy distribution. Five segmented einzel lenses<sup>53</sup> and an electrostatic double-focusing 90° deflector<sup>54</sup> are employed for focusing and steering. A pair of pulsed drift tubes (following the approach described in ref.<sup>33</sup>) is used for deceleration and pre-cooling, reducing the phase-space volume of the bunches. Downstream, two microchannel plate (MCP) detectors can be moved into the ion beam in front of and behind the Paul trap to optimize ion yield and beam transmission. The first MCP detector also features a retarding field analyser that is used to determine the mean kinetic energy and the energy spread of the ion bunches. Although this method of HCI production, transfer and recapture combining the EBIT, beamline and Paul trap can handle a large variety of elements and charge states, the following section refers specifically to the present case of optimized  $^{40}\text{Ar}^{13+}$  recapture.

Highly charged argon ions are produced in a distribution of charge states using a 13-mA, 1-keV electron beam in the approximately 50-V-deep axial trapping potential of the EBIT. In each cycle, the central trap electrode is rapidly switched from about 450 V (for the aforementioned 50-V-deep trap) to a repulsive extraction potential of 700 V (200 V higher than the outer trap electrode potential) for a period of 200 ns at a rate of 4 Hz to eject the ions. The kinetic energy relative to the ground potential of the beamline (0 V), and thus the velocity of the extracted ions, depend on the total extraction potential of 700 V and on the ionic charge  $q$ , allowing separation of the different charge states by their different times of flight (Extended Data Fig. 2a, b).  $^{40}\text{Ar}^{13+}$  is selected with the help of an electrode of the third segmented einzel lens immediately behind the 90° deflector. This electrode acts as a gate by rapidly switching to a passing voltage at the  $^{40}\text{Ar}^{13+}$  arrival time, and back to a deflecting voltage after the ion passage. Thus, the trajectories of all other charge states are deflected away from the Paul trap. We measure a mean kinetic energy of  $694q$  V with respect to ground for the fast  $^{40}\text{Ar}^{13+}$  bunches using the retarding field analyser (Extended Data Fig. 2e). An associated axial energy spread of  $32q$  V was also determined. To decrease the mean kinetic energy and its spread to values more amenable for trapping and efficient cooling in the cryogenic Paul trap, we perform an electrodynamic deceleration step with the pair of pulsed drift tubes. By biasing them to approximately 510 V and 590 V before the extraction, a linear axial potential gradient is generated on the beamline axis between the two electrodes. Thus, when the ion bunch arrives at that position, about 9.7  $\mu\text{s}$  after ion ejection, it is exposed to a mean potential of 550 V. Then, both drift tube potentials are rapidly grounded using a fast high-voltage switch. This slows down the ion bunch to a kinetic energy of  $146q$  V and reduces the axial energy spread to  $13q$  V (Extended Data Fig. 2f). The deceleration step also shortens the length of the ion bunches considerably, from about 5.2 cm to about 1.7 cm FWHM, while their temporal width is only slightly reduced (Extended Data Fig. 2c, d). After passing through a final einzel lens and an unbiased mirror tube, the  $^{40}\text{Ar}^{13+}$  ions enter the Paul trap. The trap voltages are commonly biased to 138 V to accomplish the final electrostatic deceleration step. This brings the ions to a residual kinetic energy of about  $5q$  V to  $10q$  V.

The Paul trap is formed by a radially confining radio-frequency potential and an axially confining d.c. potential. Once inside the trap, the HCIs repeatedly pass through a cigar-shaped Coulomb crystal composed of about 50 to 100  $^9\text{Be}^+$  ions that has been previously loaded into the Paul trap using laser ablation combined with photoionization<sup>34</sup> (see also Fig. 1). This proceeds as follows. During injection into the Paul

trap, owing to their relatively high kinetic energy, most HCIs can overcome the weak axially confining potential of 300 mV (above the biased ground of 138 V) applied to the electrostatic endcap at the entrance of the trap. After passing through the  $^9\text{Be}^+$  Coulomb crystal for the first time, the  $^{40}\text{Ar}^{13+}$  ions are reflected by the opposite electrostatic endcap potential of about 12 V (above the biased ground of 138 V). In the meantime (17.1  $\mu\text{s}$  after initial ion extraction from the EBIT), the mirror tube at the entrance of the Paul trap is rapidly switched up to a confining axial electrostatic potential to complete the capture of  $^{40}\text{Ar}^{13+}$ . Then the trap remains closed for 1.9 s, during which the HCIs can dissipate their residual kinetic energy by repeated interactions with the laser-cooled  $^9\text{Be}^+$  ions. If these steps are successful for an  $^{40}\text{Ar}^{13+}$  ion, it joins the  $^9\text{Be}^+$  Coulomb crystal (Fig. 1). Otherwise, the mirror-tube potential is lowered again to let the next HCI bunch enter the Paul trap. This whole recapture process is rather efficient and succeeds in less than 30 s on average.

### Excited-state lifetime measurement

The data for the lifetime measurement were acquired from 440 measurements, each of which includes 100 experimental realizations, adding up to a total measurement time of about two hours. Eleven measurements were averaged for every single wait time, with the error bars in Fig. 3c indicating the quantum projection noise of 1,100 experimental implementations. To cancel the effects of parameter drifts on the observed signal, the wait time was scanned in a pseudo-random sequence.

Drifts of the atomic resonance frequencies could lead to systematic variations in the detected excitation probabilities. The shortest achievable  $\pi$ -times, 16  $\mu\text{s}$  for the initial HCI excitation and 225  $\mu\text{s}$  for the HCI sideband transition, lead to an interaction broadening of the respective lines of 62 kHz and 4.4 kHz, respectively. Our typical short-term magnetic-field fluctuations lead to line shifts of <10 Hz level, and thus affect the measured excited-state population of the order of  $10^{-4}$ . The axial trap frequency has fluctuations below 100 Hz over the course of a day. The distribution of data points for a given wait time is consistent with the expected quantum projection noise, thereby ruling out systematic drifts at the level of the statistical uncertainty.

The clock laser pulses are generated by the first diffraction order of an acousto-optic modulator (AOM). Despite the typical 100-dB level of extinction of the radio-frequency drive power provided by an active radio-frequency switch, the optical extinction ratio does not reach this level owing to scattered light within the AOM crystal. However, this leaked light is unshifted by the AOM and therefore detuned from the ion resonance by the radio-frequency drive frequency of about 200 MHz, or  $10^7$  natural linewidths. This alone reduces the de-excitation probability by approximately 14 orders of magnitude.

Spontaneous decay of the HCI on the red sideband is suppressed as the square of the Lamb–Dicke parameter,  $\eta^2 \approx 0.01$ . However, residual decay on this sideband leads to heating of the motional mode and may thus appear as spurious excitation in the quantum logic detection. A few sideband cooling pulses applied immediately before the quantum logic transfer pulse suppress this effect by returning the crystal to its ground state. Off-resonant depumping of the excited state of  $^{40}\text{Ar}^{13+}$  by the  $^9\text{Be}^+$  lasers is negligible because of the narrow natural linewidth and the large detuning. Collisional deshelling, as discussed in refs.<sup>55,56</sup>, is absent in this experiment owing to the extremely high vacuum. Furthermore, collisions of an HCI with a neutral particle probably lead to charge exchange and total, but inconsequential, ion loss.

### g-factor evaluation

The  $^{40}\text{Ar}^{13+}2\text{P}_{3/2}$  excited-state g-factor, denoted as  $g_{3/2}$ , is determined by a linear fit of the Zeeman substate energy shifts. We use the well known ground-state g-factor of the clock transition from a recent high-accuracy measurement<sup>17</sup> to operate a co-magnetometer and measure the magnetic field by an appropriate combination of the Zeeman components.

The energy shifts  $\Delta E_{3/2, m_{3/2,i}} = m_{3/2,i} g_{3/2} \mu_B B$  of the Zeeman substates of the excited  $^2P_{3/2}$  state that are due to an external magnetic field  $B$  are obtained from the measured Zeeman shifts  $f_i$  (in units of frequency) of the six Zeeman components ( $i$  ranging from 1 to 6, according to Fig. 4c) and analogously the shifts  $\Delta E_{1/2, m_{1/2,i}}$  of the  $^2P_{1/2}$  Zeeman substates.  $h$  and  $\mu_B$  are the Planck constant and the Bohr magneton, respectively. The shifts are referenced with respect to the degenerate line/level centres. One then obtains

$$\frac{\Delta E_{3/2, m_{3/2,i}}}{h} = f_i + \frac{\Delta E_{1/2, m_{1/2,i}}}{h} \quad (1)$$

$$m_{3/2,i} g_{3/2} \frac{\mu_B B}{h} = f_i + m_{1/2,i} g_{1/2} \frac{\mu_B B}{h} \quad (2)$$

$B$  is eliminated from the above equation by using the four inner Zeeman components 2–5, which are less sensitive to magnetic-field fluctuations than the two outer ones. Components  $f_2$  and  $f_3$  share the common excited state  $m_{3/2} = -1/2$  (see Fig. 4c), and therefore their difference yields the ground-state Zeeman splitting directly, without relying on the excited-state  $g$ -factor. Using the known ground-state  $g$ -factor  $g_{1/2}$  from the work of Arapoglou et al.<sup>17</sup>, we obtain the magnetic field

$$B_1 = \frac{h(f_3 - f_2)}{g_{1/2} \mu_B} \quad (3)$$

Similarly, components  $f_4$  and  $f_5$  share the excited state with  $m_{3/2} = +1/2$ , and we acquire a second measurement of

$$B_2 = \frac{h(f_5 - f_4)}{g_{1/2} \mu_B} \quad (4)$$

Introducing  $U = f_5 - f_4 + f_3 - f_2$  for simplicity, we average the magnetic field  $B$  from these two relations to reduce the uncertainty

$$B = \frac{B_1 + B_2}{2} = \frac{h U}{2 g_{1/2} \mu_B} \quad (5)$$

This expression is inserted into equation (2) to obtain

$$y_i(m_{3/2,i}) = \frac{g_{3/2} U}{2 g_{1/2}} m_{3/2,i} = f_i + m_{1/2,i} \frac{U}{2} \quad (6)$$

On the right-hand side of the equation, the measured shifts of the excited Zeeman substates are given, which fulfil a linear relation in  $m_{3/2,i}$  (left-hand side of the equation). A linear fit (see black line in Fig. 4b) of the form

$$y_i(m_{3/2,i}) = a m_{3/2,i} + b \quad (7)$$

with offset  $b$  to account for the global frequency offset in the measured  $f_i$ , allows us to determine the excited-state  $g$ -factor  $g_{3/2}$  from the slope  $a$

$$g_{3/2} = \frac{2 g_{1/2} a}{U} \quad (8)$$

The uncertainties  $\sigma_{y_i}$  of the excited Zeeman substates are obtained from the right-hand side of equation (6) by expressing  $U$  again as  $U = f_5 - f_4 + f_3 - f_2$ , followed by standard uncertainty propagation with the independently measured  $f_i$

$$\sigma_{y_i} = \sqrt{\sum_j \left( \frac{\partial y_i}{\partial f_j} \sigma_{f_j} \right)^2} \quad (9)$$

The uncertainties  $\sigma_{f_i}$  of the Zeeman components depend on the statistical uncertainty of the line centre from the fit,  $\sigma_{f_{i,\text{fit}}}$  (fitting the lines by Rabi line shapes) and the relative systematic magnetic-field uncertainty  $\sigma_B/B$ . The latter is time-dependent and is estimated from the observed magnetic-field stability measured previously by using the  $^9\text{Be}^+$  qubit transition frequency (see ref. <sup>34</sup> for details) to be  $4.1 \times 10^{-6}$  (measurement 1) and  $3.2 \times 10^{-6}$  (measurements 2 and 3) on relevant timescales. Accordingly, one has

$$\sigma_{f_i} = \sqrt{\sigma_{f_{i,\text{fit}}}^2 + \left( \frac{\sigma_B}{B} f_i \right)^2} \quad (10)$$

The linear fit shown in Fig. 4b is weighted with the  $\sigma_{y_i}$  uncertainties, which are displayed in the lower panel. For completeness, we state the fit offsets for the three sets of measurements:  $b = -17(3)$  Hz,  $-45(2)$  Hz and  $-30(2)$  Hz. The reduced  $\chi^2$  of the linear fits are 1.45, 0.57 and 0.21.

To estimate the uncertainty of  $g_{3/2}$ , we replace  $a$  and  $U$  in equation (8)

by their analytical expressions.  $a = \overline{(m_{3/2,i} y_i)} / \overline{m_{3/2,i}^2}$  is obtained from the closed-form solution of a linear fit. The  $y_i$  values are given by the right-hand side of equation (6), and  $U = f_5 - f_4 + f_3 - f_2$ . We can neglect the parts-per-billion uncertainty of the experimental result  $g_{1/2} = 0.66364845532(93)$  from the very recent Penning trap measurement<sup>17</sup> because it is more than three orders of magnitude smaller than our experimental uncertainties in the 25,000-times-weaker magnetic field. Finally, the only uncertainties are introduced by the independently measured  $f_i$ . Thus, the uncertainty  $\sigma_{g_{3/2}}$  is obtained from the typical formula of uncertainty propagation

$$\sigma_{g_{3/2}} = \sqrt{\sum_i \left( \frac{\partial g_{3/2}}{\partial f_i} \sigma_{f_i} \right)^2} \quad (11)$$

The calculated  $g_{3/2}$  values are  $1.3322989(19)_{\text{stat}}(56)_{\text{syst}}$ ,  $1.3322897(23)_{\text{stat}}(43)_{\text{syst}}$  and  $1.3322828(24)_{\text{stat}}(43)_{\text{syst}}$  for the three measurement sets obtained on two different days, where we have stated the statistical and systematic uncertainties separately. The results are shown in Fig. 5 together with recent calculations. The uncertainties of the individual measurements are the root of the sum of the squared statistical and systematic uncertainties. The measurements agree within their uncertainties, and the largest deviation between measurement 1 and the weighted average is 1.6 standard deviations. We obtain the weighted average  $g_{3/2} = 1.3322895(13)_{\text{stat}}(56)_{\text{syst}}$ , where we have combined the statistical uncertainties and stated the largest systematic uncertainty of the individual measurements as a conservative estimate for the systematic uncertainty of the average.

## Data availability

The datasets generated and analysed during this study are available from the corresponding author upon reasonable request.

53. Mandal, P., Sikler, G. & Mukherjee, M. Simulation study and analysis of a compact einzel lens-deflector for low energy ion beam. *J. Instrum.* **6**, P02004 (2011).
54. Kreckel, H. et al. A simple double-focusing electrostatic ion beam deflector. *Rev. Sci. Instrum.* **81**, 063304 (2010).
55. Barton, P. A. et al. Measurement of the lifetime of the  $3d^2D_{3/2}$  state in  $^{40}\text{Ca}^+$ . *Phys. Rev. A* **62**, 032503 (2000).
56. Letchumanan, V., Wilson, M., Gill, P. & Sinclair, A. Lifetime measurement of the metastable  $4d^2D_{3/2}$  state in  $^{88}\text{Sr}^+$  using a single trapped ion. *Phys. Rev. A* **72**, 012509 (2005).

**Acknowledgements** We acknowledge I. Arapoglou, H. Bekker, S. Bernitt, K. Blaum, A. Egl, S. Hannig, S. Kühn, T. Legero, R. Müller, J. Nauta, J. Stark, U. Sterr, S. Sturm and A. Surzhykov for support and discussions. We also thank the MPIK engineering design office, the electronics workshops of QUEST and MPIK, IMPT Hannover, and PTB division 4 for support and technical help. In particular, we thank the mechanical workshop of MPIK and the scientific instrumentation department (5.5) of PTB for their skilful and timely manufacturing of our devices. The project was supported by the Physikalisch-Technische Bundesanstalt, the Max-

# Article

Planck Society, the Max-Planck-Riken–PTB–Center for Time, Constants and Fundamental Symmetries, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SCHM2678/5-1, the collaborative research centres SFB 1225 ISOQUANT and SFB 1227 DQ-mat, and Germany's Excellence Strategy – EXC-2123/1 QuantumFrontiers. This project also received funding from the European Metrology Programme for Innovation and Research (EMPIR), which is co-financed by the Participating States, and from the European Union's Horizon 2020 research and innovation programme (project number 17FUN07 CC4C). S.A.K. acknowledges financial support from the Alexander von Humboldt Foundation.

**Author contributions** P.M., T.L., S.A.K., E.B., L.S., M.S., J.R.C.L.-U. and P.O.S. developed the experimental setup. P.M., T.L., S.A.K. and L.J.S. carried out the experiments. P.M. and T.L.

analysed the data. J.R.C.L.-U. and P.O.S. conceived and supervised the study. P.M. and P.O.S. wrote the initial manuscript with contributions from T.L., S.A.K. and J.R.C.L.-U. All authors discussed the results and reviewed the manuscript.

**Competing interests** The authors declare no competing interests.

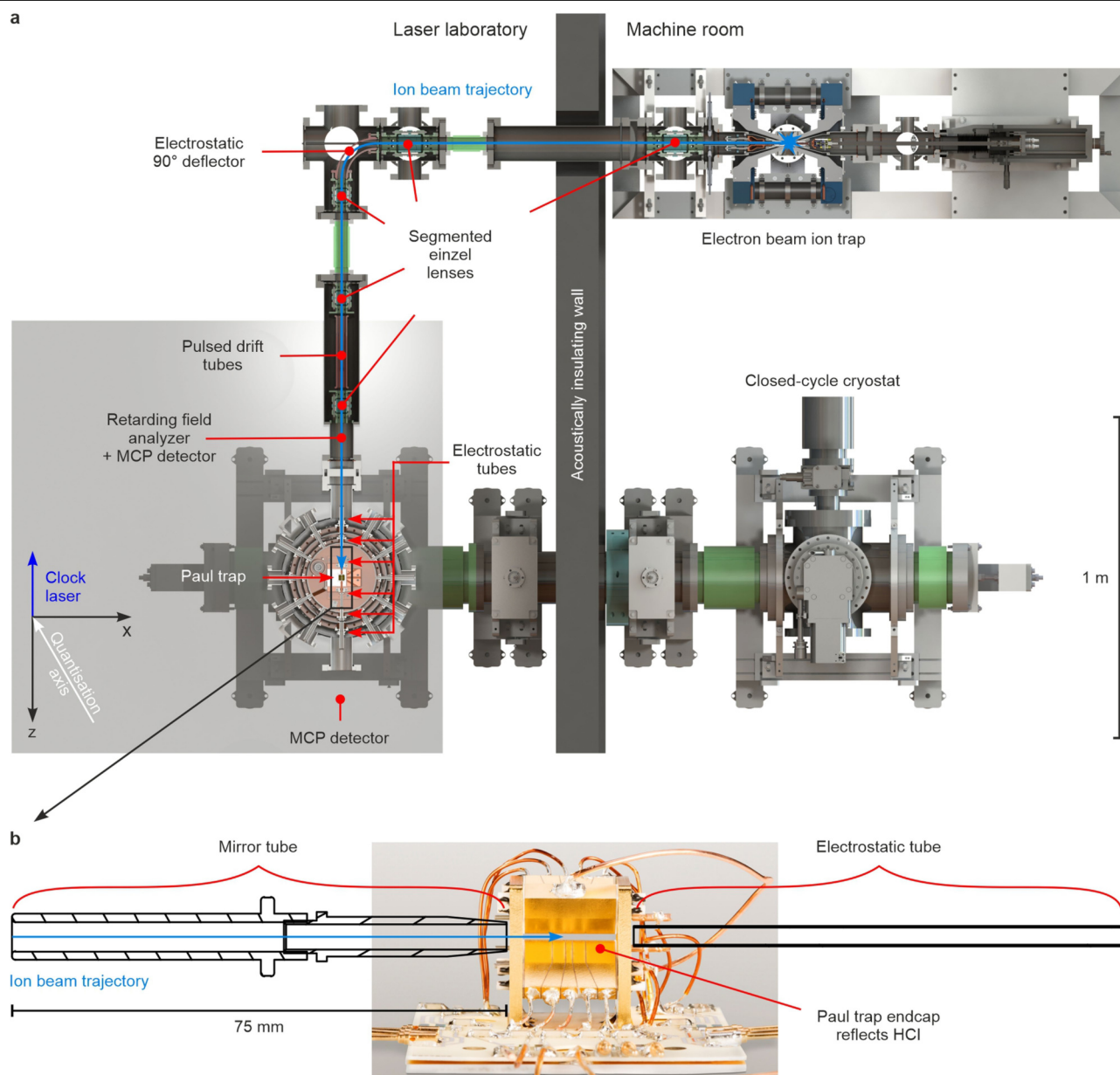
## Additional information

**Correspondence and requests for materials** should be addressed to P.M. or P.O.S.

**Peer review information** *Nature* thanks Andrei Derevianko and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

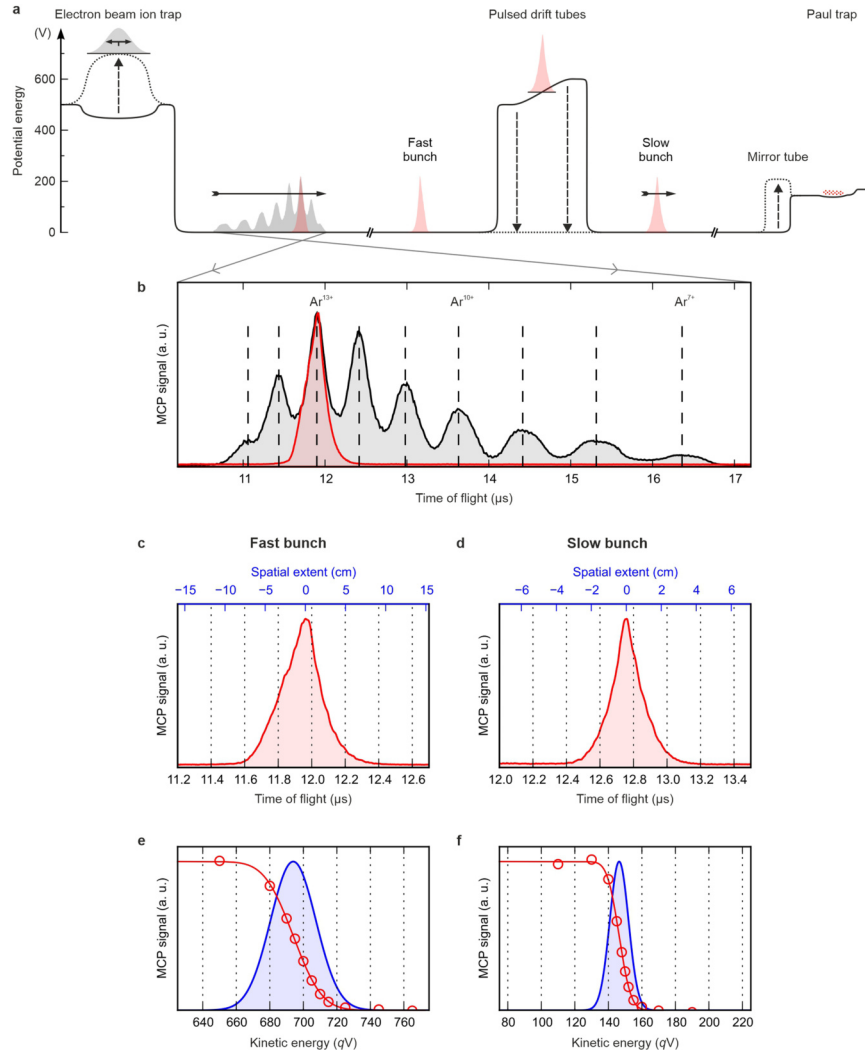
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





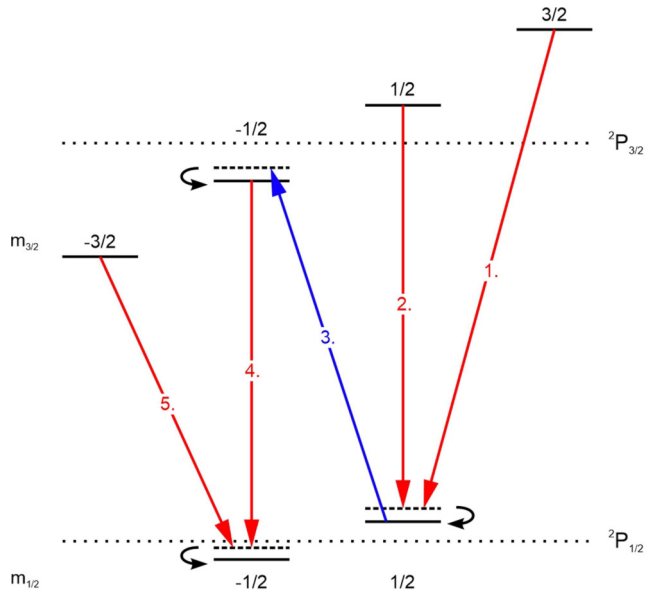
**Extended Data Fig. 1 | Experimental setup. a,** Top view of the setup. The apparatus extends over two rooms separated by an acoustically insulating wall. Inside the ‘machine room’ on the right-hand side, HCIs are produced in an EBIT<sup>32</sup> and extracted as ion bunches along the ion beam trajectory (blue line) through a deceleration beamline. At the laser laboratory (left side), they are axially injected into a cryogenic linear Paul trap<sup>34</sup>, which is mounted on a pneumatically floating optical table (grey-shaded). The Paul trap is refrigerated by a vibrationally decoupled pulse tube cryocooler<sup>35</sup> located in the machine room. The beamline is composed of several ion optical elements: five segmented einzel lenses and an electrostatic 90° deflector for guiding and focusing the ions, a pair of pulsed drift tubes for deceleration, and six cylindrical electrodes arranged in line in front of and behind the Paul trap.

Charge-state separation is accomplished by the different times of flight through the beamline. One electrode of the third segmented einzel lens is used as a gate to select the desired charge state. An MCP detector in front of the Paul trap includes two fine stainless-steel meshes that apply a well defined retarding field, and allows the measurement of the kinetic-energy distribution of the ion bunches (see also Extended Data Fig. 2e, f). A second MCP detector behind the Paul trap is used to optimize the ion beam transmission through the Paul trap. **b,** Magnified side view of the cryogenic Paul trap region. The trap (photograph) is shown with the two adjacent electrostatic tubes. The left one (mirror tube) at the entrance of the Paul trap is used to capture the HCIs by rapidly switching to a confining potential once the HCIs have passed it. Photograph: Physikalisch-Technische Bundesanstalt.



**Extended Data Fig. 2 | HCI extraction and transfer.** **a**, Simplified illustration of the electrostatic potential used for the  $^{40}\text{Ar}^{13+}$  transfer from the EBIT to the Paul trap. The entire ion inventory stored in the EBIT, with its charge-state distribution displayed as grey-shaded, is ejected by switching the axial trap to repulsive potential. The charge states separate owing to their distinct initial kinetic energies.  $^{40}\text{Ar}^{13+}$  ions (red) are selected by an electrode used as a gate (not shown). The fast  $^{40}\text{Ar}^{13+}$  bunch is then slowed down upon entering the pulsed drift tubes. Having arrived there at the centre of a linear potential gradient, the electrode potentials are rapidly switched to ground, and a slower  $^{40}\text{Ar}^{13+}$  bunch leaves the pulsed drift tubes. At the Paul trap, the ions are further decelerated by an electrostatic potential and enter the trapping region with a reduced residual kinetic energy of  $5q\text{ V}$  to  $10q\text{ V}$ . They then pass a Coulomb crystal of  $^9\text{Be}^+$  ions and are reflected by an electrostatic endcap electrode biased to a potential of about  $12\text{ V}$  above the biased common ground. Meanwhile, an electrostatic mirror tube in front of the Paul trap has been switched up to a confining potential at which  $^{40}\text{Ar}^{13+}$  is unable to escape the Paul trap. This causes an oscillatory motion along the trap axis. Through repeated

interactions with the laser-cooled  $^9\text{Be}^+$  ions,  $^{40}\text{Ar}^{13+}$  dissipates its residual kinetic energy and joins the Coulomb crystal. **b**, Normalized ion yield as a function of the time of flight after ion ejection from the EBIT, measured by the first MCP detector in front of the Paul trap. The black curve shows the entire charge-state distribution, with Ar charge states from +7 through +15. Using the gate electrode,  $^{40}\text{Ar}^{13+}$  is chosen for passage, as shown by the red curve. a.u., arbitrary units. **c, d**, Normalized  $^{40}\text{Ar}^{13+}$  bunches as a function of time and position along the beamline axis (averaged over 16 shots). The FWHM of the fast bunch is about  $250\text{ ns}$  (**c**) and that of the slow bunch is about  $185\text{ ns}$  (**d**). **e, f**, Normalized kinetic-energy distributions of the  $^{40}\text{Ar}^{13+}$  bunches along the beamline axis: fast bunch (**e**) and slow bunch after deceleration and phase-space cooling using the pulsed drift tubes (**f**). The red circles show the integrated ion yield of an averaged  $^{40}\text{Ar}^{13+}$  bunch (16 shots) for a given retardation potential, measured by the retarding-field analyser. A Gaussian error function (red line) was fitted to the data and differentiated to obtain the Gaussian energy distribution (blue line) to show the mean kinetic energy and longitudinal energy spread.



**Extended Data Fig. 3 | Quantum logic-assisted internal state preparation of  $\text{Ar}^{13+}$ .** The  $m_{1/2} = -1/2$  state of the  $^2P_{1/2}$  level is deterministically populated by a series of five clock laser sideband  $\pi$ -pulses (1–5), which excite the two-ion crystal from the motional ground state  $|0\rangle_m$  (solid lines) into the excited state  $|1\rangle_m$  (dashed lines). By means of Raman sideband cooling pulses acting on the  $^9\text{Be}^+$  ion, the crystal is returned to the motional ground state after each transfer pulse. This ensures unidirectional optical pumping<sup>9</sup>. To increase the state-preparation efficiency, this sequence is repeated four times. The other Zeeman ground state ( $^2P_{1/2}, m_{1/2} = +1/2$ ) is prepared in an analogous manner.

# Quantum crystal structure in the 250-kelvin superconducting lanthanum hydride

<https://doi.org/10.1038/s41586-020-1955-z>

Received: 24 July 2019

Accepted: 14 November 2019

Published online: 5 February 2020

Ion Errea<sup>1,2,3</sup>, Francesco Belli<sup>1,2</sup>, Lorenzo Monacelli<sup>4</sup>, Antonio Sanna<sup>5</sup>, Takashi Koretsune<sup>6</sup>, Terumasa Tadano<sup>7</sup>, Raffaello Bianco<sup>2</sup>, Matteo Calandra<sup>8</sup>, Ryotaro Arita<sup>9,10</sup>, Francesco Mauri<sup>4,11</sup> & José A. Flores-Livas<sup>4\*</sup>

The discovery of superconductivity at 200 kelvin in the hydrogen sulfide system at high pressures<sup>1</sup> demonstrated the potential of hydrogen-rich materials as high-temperature superconductors. Recent theoretical predictions of rare-earth hydrides with hydrogen cages<sup>2,3</sup> and the subsequent synthesis of LaH<sub>10</sub> with a superconducting critical temperature ( $T_c$ ) of 250 kelvin<sup>4,5</sup> have placed these materials on the verge of achieving the long-standing goal of room-temperature superconductivity. Electrical and X-ray diffraction measurements have revealed a weakly pressure-dependent  $T_c$  for LaH<sub>10</sub> between 137 and 218 gigapascals in a structure that has a face-centred cubic arrangement of lanthanum atoms<sup>5</sup>. Here we show that quantum atomic fluctuations stabilize a highly symmetrical  $Fm\bar{3}m$  crystal structure over this pressure range. The structure is consistent with experimental findings and has a very large electron–phonon coupling constant of 3.5. Although ab initio classical calculations predict that this  $Fm\bar{3}m$  structure undergoes distortion at pressures below 230 gigapascals<sup>2,3</sup>, yielding a complex energy landscape, the inclusion of quantum effects suggests that it is the true ground-state structure. The agreement between the calculated and experimental  $T_c$  values further indicates that this phase is responsible for the superconductivity observed at 250 kelvin. The relevance of quantum fluctuations calls into question many of the crystal structure predictions that have been made for hydrides within a classical approach and that currently guide the experimental quest for room-temperature superconductivity<sup>6–8</sup>. Furthermore, we find that quantum effects are crucial for the stabilization of solids with high electron–phonon coupling constants that could otherwise be destabilized by the large electron–phonon interaction<sup>9</sup>, thus reducing the pressures required for their synthesis.

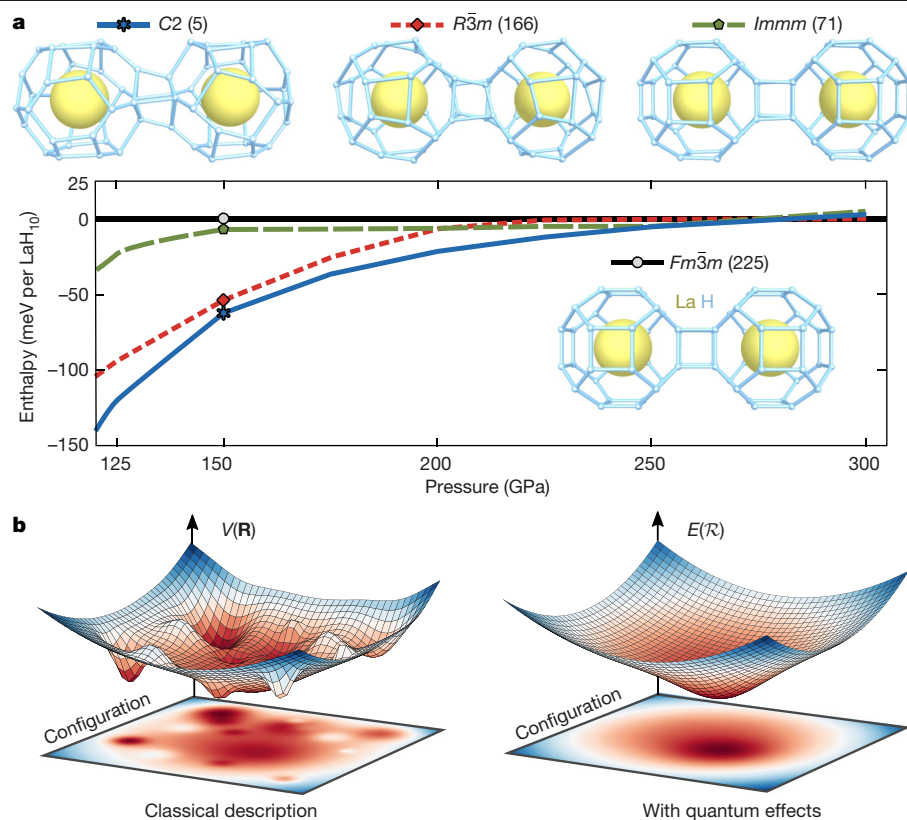
The potential of metallic hydrogen as a high- $T_c$  superconductor<sup>10</sup> was identified shortly after the development of Bardeen–Cooper–Schrieffer theory, which explains superconductivity using the electron–phonon coupling mechanism. The main argument in favour of metallic hydrogen was that  $T_c$  can be maximized for light compounds owing to their high vibrational frequencies. Because high pressures are required to metallize hydrogen<sup>11</sup>, chemical precompression with heavier atoms<sup>12,13</sup> was suggested as a pathway by which to decrease the pressure needed to reach metallicity and, therefore, superconductivity. These ideas have been realised using modern ab initio crystal structure prediction methods based on density functional theory (DFT)<sup>7,14,15</sup>. Hundreds of hydrogen-rich compounds have been predicted to be thermodynamically stable at high pressures, and their  $T_c$  values have been estimated by calculating the electron–phonon interaction<sup>6,7</sup>. The success of this co-operation between DFT crystal-structure predictions and  $T_c$  calculations was exemplified by the discovery of superconductivity

in H<sub>3</sub>S at 200 K<sup>1,16,17</sup>. The prospects for discovering hydrogen-based high- $T_c$  superconductors in the near future are therefore high, with rare-earth hydrides with sodalite-like clathrate structures showing particular promise<sup>2,3</sup>. This is in clear contrast to other high- $T_c$  superconducting families such as cuprates or pnictides, in which the lack of a clear understanding of the superconducting mechanism hinders an in silico-guided approach.

DFT predictions of the La–H system proposed LaH<sub>10</sub> to be thermodynamically stable against decomposition at high pressures<sup>2</sup>. A sodalite-type structure with space group  $Fm\bar{3}m$  and  $T_c \approx 280$  K at pressures greater than around 220 GPa was suggested as a candidate for high- $T_c$  superconductivity<sup>2,3</sup> (Fig. 1). Distorted versions of the  $Fm\bar{3}m$  structure with space group  $C2/m$  and a rhombohedral lanthanum sublattice were also discussed<sup>18</sup>, and shortly after the first predictions<sup>2,3</sup>, a lanthanum superhydride was synthesized by heating a lanthanum sample with a laser in a hydrogen-rich atmosphere inside a diamond anvil cell (DAC)<sup>19</sup>.

<sup>1</sup>Fisika Aplikatua 1 Saila, Gipuzkoako Ingeniaritza Eskola, University of the Basque Country (UPV/EHU), San Sebastián, Spain. <sup>2</sup>Centro de Física de Materiales (CSIC-UPV/EHU), San Sebastián, Spain. <sup>3</sup>Donostia International Physics Center (DIPC), San Sebastián, Spain. <sup>4</sup>Dipartimento di Fisica, Università di Roma La Sapienza, Rome, Italy. <sup>5</sup>Max-Planck Institute of Microstructure Physics, Halle, Germany. <sup>6</sup>Department of Physics, Tohoku University, Sendai, Japan. <sup>7</sup>Research Center for Magnetic and Spintronic Materials, National Institute for Materials Science, Tsukuba, Japan. <sup>8</sup>Sorbonne Université, CNRS, Institut des Nanosciences de Paris, Paris, France. <sup>9</sup>Department of Applied Physics, University of Tokyo, Tokyo, Japan. <sup>10</sup>RIKEN Center for Emergent Matter Science, Wako, Japan. <sup>11</sup>Graphene Labs, Fondazione Istituto Italiano di Tecnologia, Genoa, Italy. \*e-mail: jose.flores@uniroma1.it





**Fig. 1 | Quantum effects stabilize the symmetric  $Fm\bar{3}m$  phase of  $LaH_{10}$ .** **a**, Enthalpy as a function of pressure for different structures of  $LaH_{10}$ , neglecting zero-point energy in the calculations. Here, pressure is calculated classically from  $V(\mathbf{R})$ , neglecting quantum effects on  $V(\mathbf{R})$ . **b**, Left, a Born–Oppenheimer energy surface  $V(\mathbf{R})$ , exemplifying the presence of many local minima belonging to distorted structures.  $\mathbf{R}$  represents the positions of atoms treated classically as simple points. Right, the configurational energy surface  $E(\mathcal{R})$ , including quantum effects.  $\mathcal{R}$  represents the quantum centroid positions, which determine the centre of the ionic wave functions—that is, the average atomic positions. By including quantum effects, all phases collapse to a single phase: the highly symmetric  $Fm\bar{3}m$ .

On the basis of the unit cell volume obtained by X-ray diffraction, the hydrogen-to-lanthanum ratio was estimated to be between 9 and 12. The lanthanum atoms adopted a face-centred cubic (fcc) arrangement at pressures greater than about 160 GPa, whereas at lower pressures the lattice was rhombohedral with a lanthanum sublattice of the  $R\bar{3}m$  space group. Owing to the small X-ray cross-section of hydrogen, it is not experimentally possible to resolve the hydrogen sublattice directly. More recently, evidence of a superconducting transition at 260 K and 188 GPa was reported in a lanthanum superhydride<sup>4,20</sup>. These findings were subsequently confirmed by the measurement of a  $T_c$  of 250 K from 137 to 218 GPa in a structure with an fcc arrangement of the lanthanum atoms, suggesting a  $LaH_{10}$  stoichiometry<sup>5</sup>.

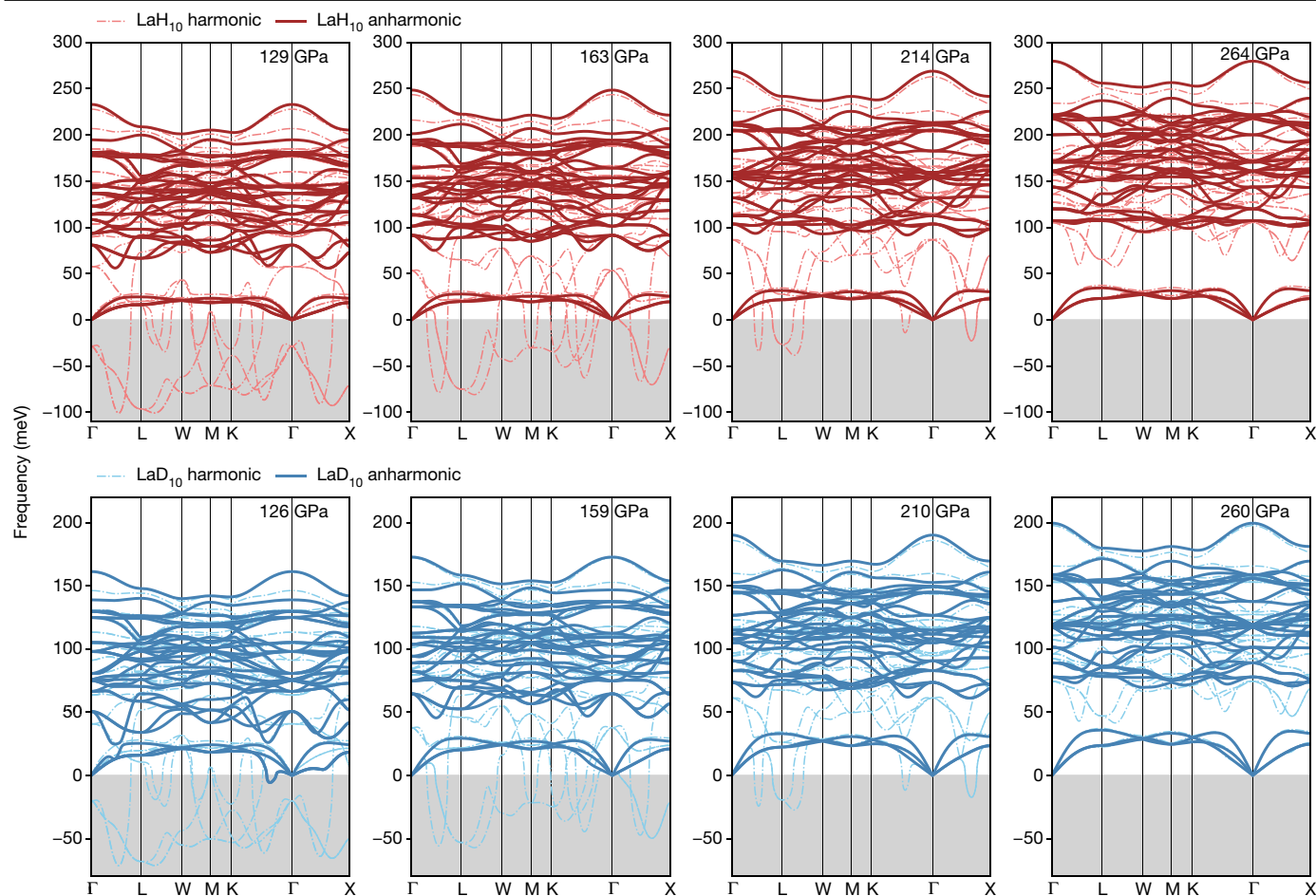
Although it is tempting to assign the superconductivity at 250 K to the previously predicted  $Fm\bar{3}m$  phase<sup>2–5</sup>, there is a clear problem: the  $Fm\bar{3}m$  structure is predicted to be dynamically unstable over the whole pressure range in which a 250 K  $T_c$  has been observed. This would imply that the  $Fm\bar{3}m$  phase is not a minimum of the Born–Oppenheimer energy surface, and consequently a  $T_c$  has not been estimated for this phase in the experimental pressure range. The contradiction between the observation of superconductivity and the predicted instability of the  $Fm\bar{3}m$  phase may indicate a problem with the classical treatment of the atomic vibrations in the calculations. Considering that quantum proton fluctuations symmetrize hydrogen bonds in the high-pressure X phase of ice<sup>21</sup> and in  $H_3S$ <sup>22,23</sup>, a similar situation is expected in  $LaH_{10}$ . Here we show how quantum atomic fluctuations completely reshape the energy landscape by removing classical local minima, rendering the  $Fm\bar{3}m$  phase the true ground state and the state responsible for the observed superconducting critical temperature.

We start by using DFT to calculate the lowest-enthalpy structures of  $LaH_{10}$  as a function of pressure, using state-of-the-art methods for the prediction of crystal structure<sup>24</sup>. The contribution associated with atomic fluctuations is not included, so that the energy corresponds solely to the Born–Oppenheimer energy  $V(\mathbf{R})$ , where  $\mathbf{R}$  represents the position of atoms treated classically as simple points. As shown in Fig. 1, different distorted phases of  $LaH_{10}$  are thermodynamically more stable

than the  $Fm\bar{3}m$  phase. At pressures greater than about 250 GPa, all phases merge to the  $Fm\bar{3}m$  symmetric phase. These results are in agreement with previous calculations<sup>2</sup>, even though we identify other possible distorted structures with lower enthalpy—such as the  $R\bar{3}m$ ,  $C2$  and  $P1$  (not shown) phases. These phases feature distortion not only in the position of the hydrogen atoms but also in the lanthanum sublattice, leading to a non-fcc arrangement that should be detectable by X-ray analysis (see Extended Data Fig. 1). The fact that many structures are predicted emphasizes that the classical  $V(\mathbf{R})$  energy surface has a multifunnel structure that is tractable to many different saddle and local minima, as shown in Fig. 1.

This picture completely changes when we include the energy of quantum atomic fluctuations—the zero-point energy. We calculate the zero-point energy within the stochastic self-consistent harmonic approximation (SSCHA)<sup>25,26</sup>. The SSCHA is a variational method that calculates the energy of the system ( $E(\mathcal{R})$ ) including atomic quantum fluctuations as a function of the centroid positions  $\mathcal{R}$ , which determine the centre of the ionic wave functions. The calculations are performed without approximating  $V(\mathbf{R})$ , keeping all of its anharmonic terms. We perform a minimization of  $E(\mathcal{R})$  and determine the centroid positions at its minimum. By calculating the stress tensor from  $E(\mathcal{R})$  (ref. 26), we relax the lattice parameters in order to find structures with isotropic stress conditions considering quantum effects. We start the quantum relaxation for both  $R\bar{3}m$  and  $C2$  phases with the lattice that yields a classical isotropic pressure of 150 GPa and vanishing classical forces—that is, calculated from  $V(\mathbf{R})$ . All quantum relaxations quickly evolve into the  $Fm\bar{3}m$  phase (Extended Data Fig. 4). This suggests that the quantum energy ( $E(\mathcal{R})$ ) landscape is much simpler than the classical  $V(\mathbf{R})$  landscape, as shown in Fig. 1, and that the ground state of  $LaH_{10}$  over the pressure range of interest is the  $Fm\bar{3}m$  phase with sodalite-type symmetry. The quantum effects are substantial, reshaping the energy landscape and stabilizing structures by more than 60 meV per  $LaH_{10}$ .

Our results further confirm that the  $Fm\bar{3}m$  phase of  $LaH_{10}$  is responsible for the superconductivity at 250 K. This is consistent with the fcc



**Fig. 2 | The phonon band structure of  $Fm\bar{3}m$   $\text{LaH}_{10}$  at different pressures.** The harmonic phonons show large instabilities in several regions of the Brillouin zone. Only at the high-pressure limit—for example, at pressures greater than 220–250 GPa—is dynamic (harmonic) stabilization reached. The anharmonic phonons obtained from the Hessian of the quantum energy  $E(\mathcal{R})$  within the SSCHA are dynamically stable over the experimentally relevant pressure range.

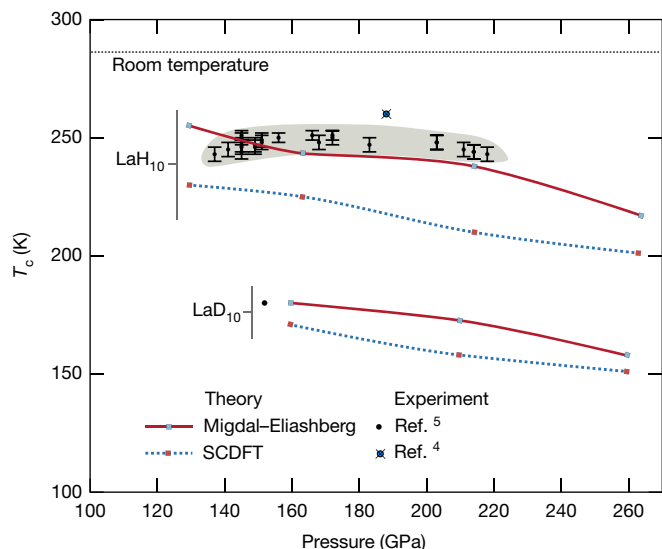
In the case of deuterium, an instability develops at low pressures (126 GPa), which is consistent with experimental evidence. The pressure stated corresponds to that calculated from  $E(\mathcal{R})$ , which considers quantum effects. The grey shading marks the regions with imaginary phonon frequencies, which are depicted as negative frequencies.

arrangement of lanthanum atoms that is found experimentally<sup>5</sup>. We verified that the experimental resolution reported<sup>5</sup> is sufficient to discard the classically obtained distorted structures (Extended Data Fig. 1). However, another study<sup>19</sup> observed a rhombohedral distortion at pressures lower than about 160 GPa, in which the lanthanum sublattice occupied the  $R\bar{3}m$  space group and the rhombohedral angle was approximately  $61.3^\circ$  ( $c/a \approx 2.38$  in the hexagonal representation). Our calculations show that this distortion is compatible with the hypothesis of slight anisotropic stress, which could be present in some experiments inside the DAC. Indeed, by performing an SSCHA minimization for the  $R\bar{3}m$  phase but keeping the rhombohedral angle fixed at  $62.3^\circ$  (the value that yields an isotropic pressure of 150 GPa at the classical level), the quantum stress tensor shows a 6% anisotropy between the diagonal direction and the perpendicular plane. This suggests that anisotropic conditions inside the DAC can produce the  $R\bar{3}m$  phase, although we cannot rule out the possibility that other experimental stress conditions could favour other crystalline phases.

The phonon spectra of the  $Fm\bar{3}m$  phase, calculated in the harmonic approximation from the Hessian of  $V(\mathbf{R})$ , show clear phonon instabilities in a broad region of the Brillouin zone (Fig. 2). These instabilities appear at pressures lower than about 230 GPa, which is consistent with the finding that many possible atomic distortions lower the enthalpy of this composition at these pressures. Conversely, when the calculation is performed using the Hessian of  $E(\mathcal{R})$  (ref. <sup>25</sup>)—which effectively

captures the full anharmonicity of  $V(\mathbf{R})$ —no phonon instability is observed (Fig. 2). This again confirms that the  $Fm\bar{3}m$  phase is a minimum in the quantum-energy landscape over the whole pressure range in which superconductivity at 250 K was observed. Whereas the  $Fm\bar{3}m$  phase of  $\text{LaH}_{10}$  remains a minimum of  $E(\mathcal{R})$  at pressures as low as 129 GPa, instabilities are seen at 126 GPa in the case of  $\text{LaD}_{10}$ . This implies that—at this pressure—the  $Fm\bar{3}m$  phase of  $\text{LaD}_{10}$  distorts to a new phase, as has been suggested previously<sup>5</sup>.

The breakdown of the classical harmonic approximation for phonons hinders the estimation of  $T_c$  at pressures lower than around 230 GPa in the  $Fm\bar{3}m$  phase. It also calls into question the certainty of harmonic calculations at higher pressures<sup>2,27</sup>, considering that large anharmonic effects are persistent well above 260 GPa (Fig. 2). However, with anharmonic phonons derived from the Hessian of  $E(\mathcal{R})$ , we can readily calculate the electron–phonon interaction and the superconducting  $T_c$  over the experimental pressure range (137–218 GPa).  $T_c$  is estimated fully ab initio—without any empirical parameter—by solving Migdal–Eliashberg equations and applying superconducting DFT (SCDFT). As shown in Fig. 3, the numerical solutions of Migdal–Eliashberg equations with an anisotropic energy gap match well with the experimental values. The values obtained from SCDFT calculations systematically show a slightly lower  $T_c$ . Our reported values of  $T_c$  provide evidence for the phonon-driven mechanism of superconductivity, and confirm that the  $Fm\bar{3}m$  structure of  $\text{LaH}_{10}$  is responsible for what is, to our knowledge,



**Fig. 3 | Summary of experimental and theoretical  $T_c$  values.** Superconducting critical temperatures calculated using anisotropic Migdal–Eliashberg equations and SCDFT. In both cases the anharmonic phonons obtained with the SSCHA were used. The results are compared with the experimental values measured in refs. <sup>4,5</sup>. The error bars in the data from ref. <sup>5</sup> correspond to the experimental uncertainty in the determination of  $T_c$ .

the highest  $T_c$  reported to date. Our calculations for  $\text{LaD}_{10}$  in the  $Fm\bar{3}m$  phase are also in agreement with the experimental data point reported. Despite the presence of large anharmonic effects, the isotope coefficient  $\alpha = -[\ln T_c(\text{LaD}_{10}) - \ln T_c(\text{LaH}_{10})]/\ln 2$  is close to 0.5 (0.43 at around 160 GPa)—as expected by Bardeen–Cooper–Schrieffer theory—and is in agreement with the experimentally reported value of  $\alpha = 0.46$ .

Finally, we also calculated  $T_c$  in the presence of the subtle rhombohedral distortion that can be induced in experiments by anisotropic pressure conditions. When the rhombohedral angle is fixed at  $62.3^\circ$ , the  $T_c$  obtained for the  $R\bar{3}m$  phase at 160 GPa is 9% lower than for the  $Fm\bar{3}m$  phase. The observed weak pressure dependence of  $T_c$  is therefore consistent with the absence of a rhombohedral distortion, as suggested by the X-ray data<sup>5</sup>. However, as argued above, undesired anisotropic stress conditions in the DAC can induce phase transitions. For cases in which measurements of  $T_c$  have yielded lower values (around 200 K), it is highly probable that the corresponding structures are distorted as a result of anisotropic pressure conditions. We can also safely rule out the possibility that compositions such as  $\text{LaH}_{11}$ —which are proposed to have a high superconducting critical temperature<sup>5</sup>—are responsible for  $T_c$  values in the range observed here (Extended Data Fig. 8).

In summary, this work demonstrates the importance of quantum effects in determining the ground-state structures of superconducting hydrides—challenging current predictions and evidencing flaws in standard theoretical methods. Similar effects are expected in other high- $T_c$  compounds with hydrogen clathrate structures, for which syntheses have recently been reported<sup>28–30</sup>. We also illustrated that quantum fluctuations are essential in order to sustain crystals with large electron–phonon coupling constants (the value of 3.6, found at 129 GPa for  $\text{LaH}_{10}$ , is to our knowledge the highest reported); such structures would otherwise be destabilized by the substantial electron–phonon interaction, resulting in distorted (low-symmetry) structures in which the electronic density of states at the Fermi level is reduced<sup>9</sup> (Extended Data Fig. 7). Our results may therefore help to increase the prospect of attaining high- $T_c$  superconductivity in hydrogen-based structures at much lower pressures than would be expected classically.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1955-z>.

1. Drozdov, A. P., Eremets, M. I., Troyan, I. A., Ksenofontov, V. & Shylin, S. I. Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system. *Nature* **525**, 73–76 (2015).
2. Liu, H., Naumov, I. I., Hoffmann, R., Ashcroft, N. W. & Hemley, R. J. Potential high- $T_c$  superconducting lanthanum and yttrium hydrides at high pressure. *Proc. Natl Acad. Sci. USA* **114**, 6990–6995 (2017).
3. Peng, F., Sun, Y., Pickard, C. J., Needs, R. J., Wu, Q. & Ma, Y. Hydrogen clathrate structures in rare earth hydrides at high pressures: possible route to room-temperature superconductivity. *Phys. Rev. Lett.* **119**, 107001 (2017).
4. Somayazulu, M. et al. Evidence for superconductivity above 260 K in lanthanum superhydride at megabar pressures. *Phys. Rev. Lett.* **122**, 027001 (2019).
5. Drozdov, A. P. et al. Superconductivity at 250 K in lanthanum hydride under high pressures. *Nature* **569**, 528–531 (2019).
6. Bi, T., Zarifi, N., Terpstra, T. & Zurek, E. The search for superconductivity in high pressure hydrides. Preprint at <https://arxiv.org/abs/1806.00163> (2018).
7. Flores-Livas, J. A., Boeri, L., Sanna, A., Profeta, G., Arita, R. & Eremets, M. A perspective on conventional high-temperature superconductors at high pressure: methods and materials. Preprint at <https://arxiv.org/abs/1905.06693> (2019).
8. Pickard, C. J., Errea, I. & Eremets, M. I. Superconducting hydrides under pressure. *Annu. Rev. Condens. Matter Phys.* <https://doi.org/10.1146/annurev-conmatphys-031218-013413> (2019).
9. Allen, P. B. & Cohen, M. L. Superconductivity and phonon softening. *Phys. Rev. Lett.* **29**, 1593 (1972).
10. Ashcroft, N. Metallic hydrogen: A high-temperature superconductor? *Phys. Rev. Lett.* **21**, 1748 (1968).
11. Dias, R. P. & Silvera, I. F. Observation of the Wigner–Huntington transition to metallic hydrogen. *Science* **355**, 715–718 (2017).
12. Gilman, J. J. Lithium dihydride fluoride—an approach to metallic hydrogen. *Phys. Rev. Lett.* **26**, 546 (1971).
13. Ashcroft, N. W. Hydrogen dominant metallic alloys: high temperature superconductors? *Phys. Rev. Lett.* **92**, 187002 (2004).
14. Zhang, L., Wang, Y., Lv, J. & Ma, Y. Materials discovery at high pressures. *Nat. Rev. Mater.* **2**, 17005 (2017).
15. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **4**, 331–348 (2019).
16. Li, Y., Hao, J., Liu, H., Li, Y. & Ma, Y. The metallization and superconductivity of dense hydrogen sulfide. *J. Chem. Phys.* **140**, 174712 (2014).
17. Duan, D. et al. Pressure-induced metallization of dense  $(\text{H}_2\text{S})_2\text{H}_2$  with high- $T_c$  superconductivity. *Sci. Rep.* **4**, 6968 (2014).
18. Liu, H., Naumov, I. I., Geballe, Z. M., Somayazulu, M., Tse, J. S. & Hemley, R. J. Dynamics and superconductivity in compressed lanthanum superhydride. *Phys. Rev. B* **98**, 100102(R) (2018).
19. Geballe, Z. M. et al. Synthesis and stability of lanthanum superhydrides. *Angew. Chem. Int. Ed.* **57**, 688–692 (2018).
20. Hemley, R. J., Ahart, M., Liu, H. & Somayazulu, M. Road to room-temperature superconductivity:  $T_c$  above 260 K in lanthanum superhydride under pressure. Preprint at <https://arxiv.org/abs/1906.03462> (2019).
21. Benoit, M., Marx, D. & Parrinello, M. Tunneling and zero-point motion in high-pressure ice. *Nature* **392**, 258–261 (1998).
22. Errea, I. et al. Quantum hydrogen-bond symmetrization in the superconducting hydrogen sulfide system. *Nature* **532**, 81–84 (2016).
23. Bianco, R., Errea, I., Calandra, M. & Mauri, F. High-pressure phase diagram of hydrogen and deuterium sulfides from first principles: structural and vibrational properties including quantum and anharmonic effects. *Phys. Rev. B* **97**, 214101 (2018).
24. Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120**, 9911 (2004).
25. Bianco, R., Errea, I., Paulatto, L., Calandra, M. & Mauri, F. Second-order structural phase transitions, free energy curvature, and temperature-dependent anharmonic phonons in the self-consistent harmonic approximation: theory and stochastic implementation. *Phys. Rev. B* **96**, 014111 (2017).
26. Monacelli, L., Errea, I., Calandra, M. & Mauri, F. Pressure and stress tensor of complex anharmonic crystals within the stochastic self-consistent harmonic approximation. *Phys. Rev. B* **98**, 024106 (2018).
27. Liu, L., Wang, C., Yi, S., Kim, K. W., Kim, J. & Cho, J.-H. Microscopic mechanism of room-temperature superconductivity in compressed  $\text{LaH}_{10}$ . *Phys. Rev. B* **99**, 140501 (2019).
28. Troyan, I. A. et al. Synthesis and superconductivity of yttrium hexahydride  $\text{Im}\bar{3}m\text{-YH}_6$ . Preprint at <https://arxiv.org/abs/1908.01534> (2019).
29. Semenok, D. V. et al. Superconductivity at 161 K in thorium hydride  $\text{ThH}_{10}$ : synthesis and properties. Preprint at <https://arxiv.org/abs/1902.10206> (2019).
30. Kong, P. P. et al. Superconductivity up to 243 K in yttrium hydrides under high pressure. Preprint at <https://arxiv.org/abs/1909.10482> (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



### Calculation details

First-principles calculations were performed within DFT and the generalized gradient approximation (GGA) as parametrized by Perdew, Burke and Ernzerhof (PBE)<sup>31</sup>. Harmonic phonon frequencies were calculated within density functional perturbation theory (DFPT)<sup>32</sup> making use of the Quantum ESPRESSO code<sup>33,34</sup>. The SSCHA<sup>25,26,35,36</sup> minimization requires the calculation of energies, forces and stress tensors in supercells. These were also calculated within DFT at the PBE level with Quantum ESPRESSO. For the final SSCHA populations, 1,000 configurations were used to reduce the stochastic noise. In all calculations we used ultrasoft pseudopotentials including 11 electrons for the La atoms, a plane-wave cut-off energy of 50 Ry for the kinetic energy and 500 Ry for the charge density.

In the harmonic phonon calculations for the  $Fm\bar{3}m$  and the  $R\bar{3}m$  phases, we used the primitive and rhombohedral lattices, respectively, with one  $LaH_{10}$  formula unit in the unit cell. A  $20 \times 20 \times 20$  Monkhorst–Pack shifted electron-momentum grid was used for these calculations with a Methfessel–Paxton smearing of 0.02 Ry. The DFT calculations performed for the SSCHA on supercells were performed on a coarser electron-momentum grid, which would correspond to a  $12 \times 12 \times 12$  grid in the unit cell. We explicitly verified that this coarser mesh yields a fully converged SSCHA gradient with respect to the electron-momentum grid, thus not affecting the SSCHA minimization. The DFT supercell calculations for the SSCHA minimization on the C2 phase were performed keeping the same  $\mathbf{k}$ -point density.

All phonon frequencies for  $\mathbf{q}$ -points that were not commensurate with the supercell used in the SSCHA minimization were obtained by directly Fourier-interpolating the real space force constants obtained in this supercell, which are calculated from the Hessian of  $E(\mathcal{R})$ . For the  $Fm\bar{3}m$  phase, the SSCHA calculation was performed both on a  $2 \times 2 \times 2$  and on a  $3 \times 3 \times 3$  supercell containing, respectively, 88 and 297 atoms. The phonon spectra shown in Fig. 2 for the  $Fm\bar{3}m$  phase were obtained by directly Fourier-interpolating the SSCHA energy Hessian force constants obtained in a  $3 \times 3 \times 3$  supercell. In Extended Data Fig. 2 we show that the phonon spectrum obtained by directly interpolating the force constants in a  $2 \times 2 \times 2$  supercell yields similar results, indicating that the energy Hessian force constants are short-range and can be Fourier-interpolated. Indeed, the  $T_c$  calculated with the  $2 \times 2 \times 2$  and  $3 \times 3 \times 3$  force constants for interpolating phonons differs by only around 3 K. Because the value of estimated  $T_c$  only negligibly depends on the cell size for the  $Fm\bar{3}m$  phase, the SSCHA quantum structural relaxations in the  $R\bar{3}m$  and C2 phases were performed in  $2 \times 2 \times 2$  supercells with 88 atoms.

As shown in ref.<sup>25</sup>, the Hessian of  $E(\mathcal{R})$  is

$$\frac{\partial^2 E(\mathcal{R})}{\partial \mathcal{R} \partial \mathcal{R}} = \Phi + \Phi \Lambda \left( 1 - \Phi \Lambda \right)^{-1} \Phi \quad (1)$$

Bold notation represents matrices and tensors in compact notation. In equation (1),  $\Phi$  are the variational force constants of the SSCHA minimization,  $\Phi^{(n)}$  the quantum statistical averages taken with the SSCHA density matrix of the  $n$ th order derivatives of  $V(\mathbf{R})$ , and  $\Lambda$  a tensor that depends on the temperature and  $\Phi$ .  $\mathbf{1}$  is the identity matrix. As we show in Extended Data Fig. 2, setting  $\Phi = 0$  has a negligible effect on the phonons obtained from the Hessian defined in equation (1). Therefore,  $\Phi$  is neglected throughout, and all superconductivity calculations in the  $Fm\bar{3}m$  and  $R\bar{3}m$  phases are performed making use of the phonon frequencies and polarization vectors obtained from the Hessian of  $E(\mathcal{R})$  with  $\Phi = 0$ . We also estimated  $T_c$  with the phonon frequencies and polarization vectors obtained instead from  $\Phi$ , resulting in a critical temperature 12 K lower within the Allen–Dynes–modified McMillan

equation. This difference is small and within the uncertainty of the  $T_c$  calculation between SCDFT and anisotropic Migdal–Eliashberg calculations (see Fig. 3 and below).

The Eliashberg spectral function, which we used for the  $T_c$  calculations, is defined as

$$\alpha^2 F(\omega) = \frac{1}{N_{E_F}} \sum_{n\mathbf{k}, m\mathbf{q}, \nu} \left| g_{n\mathbf{k}, m\mathbf{q}+\mathbf{q}}^{\nu} \right|^2 \delta(\varepsilon_{n\mathbf{k}} - E_F) \times \delta(\varepsilon_{m\mathbf{k}+\mathbf{q}} - E_F) \delta(\omega - \omega_{\mathbf{q}\nu}) \quad (2)$$

where  $N_{E_F}$  is the electronic density of states (DOS) at the Fermi energy ( $E_F$ ),  $n$  and  $m$  are band indices,  $\mathbf{k}$  is a crystal momentum,  $\varepsilon_{n\mathbf{k}}$  is a band energy,  $\omega_{\mathbf{q}\nu}$  is the phonon frequency of mode  $\nu$  at wavevector  $\mathbf{q}$ , and  $g_{n\mathbf{k}, m\mathbf{q}+\mathbf{q}}^{\nu}$  is the electron–phonon matrix element between a state  $n\mathbf{k}$  and  $m\mathbf{k} + \mathbf{q}$ . We calculated  $\alpha^2 F(\omega)$  combining the SSCHA phonon frequencies and polarization vectors obtained from the Hessian of  $E(\mathcal{R})$  with the electron–phonon matrix elements calculated with DFPT. For the  $Fm\bar{3}m$  and  $R\bar{3}m$  phases, the electron–phonon matrix elements were calculated in a  $6 \times 6 \times 6$   $\mathbf{q}$ -point grid and a  $40 \times 40 \times 40$   $\mathbf{k}$ -point grid. These were combined with the SSCHA phonons and polarization vectors obtained by Fourier interpolation to the  $6 \times 6 \times 6$   $\mathbf{q}$ -point grid from the real space force constants coming from the Hessian of  $E(\mathcal{R})$  in a  $3 \times 3 \times 3$  supercell for the  $Fm\bar{3}m$  phase and in a  $2 \times 2 \times 2$  supercell for the  $R\bar{3}m$  phase. The Dirac deltas on the band energies are estimated by substituting them with a Gaussian of width 0.004 Ry. The calculated  $\alpha^2 F(\omega)$  functions for the  $Fm\bar{3}m$  phase are shown in Extended Data Fig. 3, and in Extended Data Fig. 5 we show the results for the  $R\bar{3}m$  phase.

### Crystal phase diagram exploration

To sample the enthalpy landscape of  $LaH_{10}$  we used the minima-hopping method<sup>24,37</sup>, which has been successfully employed for global geometry optimization in a large variety of applications—including superconducting materials such as  $H_3S$ ,  $PH_3$ , and elemental solids at high pressure<sup>38–40</sup>. This composition was thoroughly explored with 1, 2, 3 and 4 formula unit simulation cells. Variable composition simulations were also performed for other La–H compositions. Energy, atomic forces and stresses were evaluated at the DFT level with the GGA-PBE parametrization to the exchange–correlation functional. A plane wave basis-set with a high cut-off energy of 900 eV was used to expand the wave function together with the projector-augmented wave method as implemented in the Vienna Ab initio Simulation Package (VASP)<sup>41</sup>. Geometry relaxations were performed with tight convergence criteria such that the forces on the atoms were less than  $2 \text{ meV } \text{\AA}^{-1}$  and the stresses were less than  $0.1 \text{ eV } \text{\AA}^{-3}$ . Extended Data Fig. 1 shows our calculated convex hull of enthalpy formation without considering the zero-point energy at 100, 150 and 200 GPa. Notably, there are many stable compositions in the convex hull.  $LaH_{10}$  becomes enthalpically stable (classically) at around 175 GPa and remains in the convex well above 300 GPa. We have verified that  $LaH_{10}$  ( $Fm\bar{3}m$ ) does not decompose into  $LaH_3$  ( $Cmcm$ ) and  $LaH_{11}$  ( $P4/nmm$ ) by around 0.3 eV per  $LaH_{10}$  once quantum effects are included in the calculation of the enthalpy formation at 150 GPa, contrary to the conclusion drawn by classical calculations in ref.<sup>3</sup>. Below 150 GPa,  $R\bar{3}m$  and C2 phases ( $LaH_{10}$ ) show unstable harmonic phonons at  $\Gamma$ , becoming saddle points of  $V(\mathbf{R})$ . However, harmonically one can find  $P1$  stable structures (decreasing symmetry) by following the instability pattern (softening direction—that is, along eigenvector polarization).  $P1$  structures are degenerate in enthalpy within less than 3 meV per  $LaH_{10}$  with respect to C2. We therefore used the C2 as a representative of highly distorted structures for our study.

### Superconductivity calculations in the $Fm\bar{3}m$ phase

Superconductivity calculations were performed within two different approaches that represent the state-of-the-art of ab initio superconductivity: SCDFT and the Eliashberg equations with full Coulomb interaction.



SCDFT is an extension to DFT for a superconducting ground state<sup>42,43</sup>. By assuming that the  $\mathbf{nk}$  anisotropy in the electron–phonon coupling is negligible (see ref. <sup>43</sup> for further details), the critical temperature is computed by solving an (isotropic) equation for the Kohn–Sham gap:

$$\Delta_s(\varepsilon) = \mathcal{Z}(\varepsilon)\Delta_s(\varepsilon) - \int d\varepsilon' \mathcal{K}(\varepsilon, \varepsilon') \frac{\tanh\left[\frac{\beta E(\varepsilon')}{2}\right]}{2E(\varepsilon')} \Delta_s(\varepsilon') \quad (3)$$

where  $\varepsilon$  is the electron energy and  $\beta$  the inverse temperature. The kernels  $\mathcal{K}$  and  $\mathcal{Z}$  come from the exchange correlation functional of the theory<sup>43–48</sup> and depend on the properties of the pairing interactions: electron–phonon coupling and screened electron–electron repulsion. Equation (3) enables us to calculate  $T_c$  completely ab initio, without introducing an empirical  $\mu^*$  parameter (Coulomb pseudopotential). Dynamic effects on the Coulomb interaction (plasmon) were also tested and did not show any substantial effect. In its isotropic form, the screened Coulomb interaction in SCDFT is accounted for by a function  $\mu(\varepsilon, \varepsilon')$ , which is given by the average<sup>49</sup> random phase approximation (RPA) Coulomb matrix element on the iso-energy surfaces  $\varepsilon$  and  $\varepsilon'$  times the DOS at  $\varepsilon'$  ( $N(\varepsilon')$ ):

$$\mu(\varepsilon, \varepsilon') = \sum_{n,m} \iint d^3(kk') V_{nk,mk'}^{\text{RPA}} \frac{\delta(\varepsilon - \varepsilon_{nk})}{N(\varepsilon_{nk})} \delta(\varepsilon' - \varepsilon_{mk'}) \quad (4)$$

The full energy dependence of the DOS is accounted for in the calculations, whereas the electron–phonon coupling is described by the  $\alpha^2 F(\omega)$  of equation (2).

The second approach we use to simulate the superconducting state is the anisotropic Eliashberg approach<sup>50</sup>. Here we include, together with the energy dependence of the electron DOS, the anisotropy of the electron–phonon coupling. The Green’s function form of the Eliashberg equation we solve is given as

$$\Sigma_{nk}(i\omega_i) = -\frac{1}{N\beta} \sum_{\mu, \mathbf{q}, m} V_{mn}^{\text{ph}}(\mathbf{q}, i\omega_\mu) G_{m\mathbf{k}+\mathbf{q}}(i\omega_\mu + i\omega_i) \quad (5)$$

$$\Delta_{nk}(i\omega_i) = -\frac{1}{N\beta} \sum_{\mu, \mathbf{q}, m} \{V_{mn}^{\text{ph}}(\mathbf{q}, i\omega_\mu) + V_{mn}^{\text{C}}(\mathbf{q}, i\omega_\mu)\} \times |G_{m\mathbf{k}+\mathbf{q}}(i\omega_\mu + i\omega_i)|^2 \Delta_{m\mathbf{k}+\mathbf{q}}(i\omega_\mu + i\omega_i) \quad (6)$$

Here,  $\Sigma_{nk}(i\omega_i)$  and  $\Delta_{nk}(i\omega_i)$  are the normal and anomalous self energy, and  $V_{mn}^{\text{ph}}(\mathbf{q}, i\omega_\mu)$  and  $V_{mn}^{\text{C}}(\mathbf{q}, i\omega_\mu)$  are the  $\mathbf{k}$ -averaged phonon-mediated interaction and Coulomb interaction, respectively. The explicit form of  $V_{mn}^{\text{ph}}(\mathbf{q}, i\omega_\mu)$  is given as

$$V_{mn}^{\text{ph}}(\mathbf{q}, i\omega_\mu) = \sum_{\mathbf{v}} |g_{nm}^{\mathbf{v}}(\mathbf{q})|^2 D_{\mathbf{v}}(\mathbf{q}, i\omega_\mu) \quad (7)$$

where  $|g_{nm}^{\mathbf{v}}(\mathbf{q})|^2$  is a  $\mathbf{k}$ -averaged electron–phonon matrix element

$$|g_{nm}^{\mathbf{v}}(\mathbf{q})|^2 = \frac{\sum_{\mathbf{k}} |g_{nk,mk+\mathbf{q}}^{\mathbf{v}}|^2 \delta(\varepsilon_{nk} - E_F) \delta(\varepsilon_{m\mathbf{k}+\mathbf{q}} - E_F)}{\sum_{\mathbf{k}} \delta(\varepsilon_{nk} - E_F) \delta(\varepsilon_{m\mathbf{k}+\mathbf{q}} - E_F)} \quad (8)$$

and  $D_{\mathbf{v}}(\mathbf{q}, i\omega_\mu)$  is a free-phonon Green’s function,  $D_{\mathbf{v}}(\mathbf{q}, i\omega_\mu) = -2\omega_{\mathbf{q}\mathbf{v}}/(\omega_\mu^2 + \omega_{\mathbf{q}\mathbf{v}}^2)$ . The electron–phonon matrix elements are calculated through a DFPT calculation with  $6 \times 6 \times 6$   $\mathbf{q}$ -point grid and are combined with the phonon frequencies and polarization vectors obtained by directly Fourier-interpolating to this grid the force constants arising from the  $E(\mathcal{R})$  Hessian in the  $3 \times 3 \times 3$  supercell. For the Coulomb interaction,  $V_{mn}^{\text{C}}(\mathbf{q}, i\omega_\mu)$  is approximated by  $\mathbf{k}$ -averaged static Coulomb interaction within the random phase approximation,  $\frac{1}{N_k} \sum_{\mathbf{k}} V_{m\mathbf{k},n\mathbf{k}+\mathbf{q}}^{\text{RPA}}(i\omega_\mu = 0)$ . Using equation (5), the Dyson equation was solved self-consistently and then equation (6) was solved to estimate  $T_c$  with  $36 \times 36 \times 36$   $\mathbf{k}$ -point grid and 512 Matsubara frequencies.

In Extended Data Table 1 we summarize all calculated  $T_c$  values within anisotropic ME and isotropic SCDFT. We also include the values obtained with the McMillan equation and the Allen–Dynes-modified McMillan equation ( $\mu^* = 0.1$ ). The calculated electron–phonon coupling constant,  $\lambda = 2 \int_0^\infty d\omega \alpha^2 F(\omega)/\omega$  and the logarithmic frequency average,  $\omega_{\log} = \exp\left(\frac{2}{\lambda} \int_0^\infty d\omega \frac{\alpha^2 F(\omega)}{\omega} \log \omega\right)$ , are also included in the table.

### Quantum structural relaxations in the $R\bar{3}m$ and $C2$ phases

In Extended Data Fig. 5 we show the evolution of the pressure calculated along the different Cartesian directions for the  $R\bar{3}m$  throughout the SSCHA minimization but keeping the rhombohedral angle fixed at  $62.3^\circ$ . Thus, the centroid positions  $\mathcal{R}$  are optimized only considering the internal degrees of freedom of the  $R\bar{3}m$  phase. Even if, at the classical level, the stress is isotropic (within a 0.5%), after the SSCHA quantum relaxation an anisotropic stress of a 6% is created between the  $z$  and  $x$ – $y$  directions. The phonons obtained at the end of the minimization are shown in Extended Data Fig. 5. Second, in Extended Data Fig. 4, we show that starting from the result of this minimization but now also relaxing the lattice, the  $R\bar{3}m$  phase evolves into the  $Fm\bar{3}m$  phase. It is clear how the pressure calculated with quantum effects becomes isotropic when the rhombohedral angle becomes  $60^\circ$ , the angle corresponding to an fcc lattice in a rhombohedral description. It is also evident that the Wyckoff positions of the  $R\bar{3}m$  phase evolve clearly into the  $Fm\bar{3}m$  Wyckoff positions, which are summarized in Extended Data Table 2.

In Extended Data Fig. 6 we show the evolution of the diagonal components of the pressure along the three different Cartesian directions for the monoclinic  $C2$  when the lattice structure is relaxed with the SSCHA. The starting point is obtained by first performing a SSCHA relaxation of only internal atomic coordinates, keeping the lattice parameters that yield an isotropic stress of 150 GPa. It is clear that quantum effects create an anisotropic stress if the lattice parameters are not modified. When the quantum relaxation of the lattice is performed, the lattice parameters are modified and an isotropic stress is recovered.

Extended Data Fig. 7 shows the structures of the  $R\bar{3}m$  and  $C2$  phases obtained classically and after the quantum SSCHA relaxation. After the quantum relaxation, the symmetry of both structures is recognized as  $Fm\bar{3}m$  with a tolerance of 0.001 Å for lattice vectors and 0.005 Å for ionic positions, consistent with the stochastic accuracy of the SSCHA. In the same figure, the electronic DOS as a function of pressure is plotted. A highly symmetric motif ( $Fm\bar{3}m$ ) maximizes  $N_{E_F}$ , whereas in distorted structures ( $R\bar{3}m$  and  $C2$ ) the occupation at the Fermi level is reduced by more than 20%. This underlines that the classical distortions would lower  $N_{E_F}$ , reducing  $\lambda$ , as expected in a system that is destabilized by the electron–phonon interaction.

### Transition temperatures from other La–H compositions

Different compositions on the La–H phase diagram have been reported to be thermodynamically stable. Presumably, the stabilization of these compositions and the measurement of different  $T_c$  values (see ref. <sup>5</sup>) demonstrate that other stoichiometries are responsible for these measured  $T_c$  values. Notably, these  $T_c$  values appear substantially lower—for instance, the values decrease from 250 K, to 215 K, 110 K and to 70 K. Experimentally there is not a clear correlation between sample preparation,  $T_c$  and pressure. In the sample preparation in ref. <sup>5</sup>, pressures can vary from 100 to 200 GPa (gradient inside the DAC) and it was proposed that other stoichiometries (low-hydrogen content) are responsible for systematically lower values of  $T_c$ .

Conversely, in a later publication, the same authors suggested that another hydrogen-rich system that is enthalpically competitive ( $\text{LaH}_{11}$ ) could possibly be responsible for other high- $T_c$  values. In order to verify this hypothesis, we considered structure-prediction runs with this stoichiometry, and found crystalline structures that were previously reported in ref. <sup>3</sup>. Extended Data Fig. 8 shows the structural motif and

the corresponding phonons and  $\alpha^2F(\omega)$  spectral function. We can rule out the possibility that high  $T_c$  values, as measured in different samples, arise from  $\text{LaH}_{11}$  in its  $P4/nmm$  (129) structure (lowest enthalpy structure for this composition at relevant experimental pressures). As seen in Extended Data Fig. 8, this phase has a strong molecular-crystal character, composed of  $\text{H}_2$  units weakly interacting with the La lattice. This phase is indeed a poor metal—with low occupation of electrons at the Fermi level—owing to its molecular character, and it cannot explain  $T_c$  values of 70 K or higher. Our estimated  $T_c$  with the Allen–Dynes formula, harmonic phonons and using a  $\mu^* = 0.1$  is 7 K at 100 GPa, reaching around 24 K at 200 GPa. More importantly, this phase does not show marked anharmonicity.

## Data availability

All the data generated in this work is available upon request from I.E. and J.A.F.-L.

## Code availability

Quantum ESPRESSO is an open-source suite of computational tools available at <https://www.quantum-espresso.org>. VASP is a proprietary program. The SSCHA and the SCDFT codes are private codes developed by some of the authors, and are being prepared for distribution as an open-source code.

31. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
32. Baroni, S., de Gironcoli, S., Dal Corso, A. & Giannozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73**, 515 (2001).
33. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
34. Giannozzi, P. et al. Advanced capabilities for materials modelling with QUANTUM ESPRESSO. *J. Phys. Condens. Matter* **29**, 465901 (2017).
35. Errea, I., Calandra, M. & Mauri, F. First-principles theory of anharmonicity and the inverse isotope effect in superconducting palladium-hydride compounds. *Phys. Rev. Lett.* **111**, 177002 (2013).
36. Errea, I., Calandra, M. & Mauri, F. Anharmonic free energies and phonon dispersions from the stochastic self-consistent harmonic approximation: application to platinum and palladium hydrides. *Phys. Rev. B* **89**, 064302 (2014).
37. Amsler, M. & Goedecker, S. Crystal structure prediction using the minima hopping method. *J. Chem. Phys.* **133**, 224104 (2010).

38. Flores-Livas, J. A., Sanna, A. & Gross, E. K. U. High temperature superconductivity in sulfur and selenium hydrides at high pressure. *Eur. Phys. J. B* **89**, 63 (2016).
39. Flores-Livas, J. A. et al. Superconductivity in metastable phases of phosphorus-hydride compounds under high pressure. *Phys. Rev. B* **93**, 020508 (2016).
40. Flores-Livas, J. A. et al. Interplay between structure and superconductivity: metastable phases of phosphorus under pressure. *Phys. Rev. Mater.* **1**, 024802 (2017).
41. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
42. Oliveira, L. N., Gross, E. K. U. & Kohn, W. Density-functional theory for superconductors. *Phys. Rev. Lett.* **60**, 2430 (1988).
43. Lüders, M. et al. Ab initio theory of superconductivity. I. Density functional formalism and approximate functionals. *Phys. Rev. B* **72**, 024545 (2005).
44. Flores-Livas, J. A. & Sanna, A. Superconductivity in intercalated group-IV honeycomb structures. *Phys. Rev. B* **91**, 054508 (2015).
45. Pellegrini, C., Glawe, H. & Sanna, A. Density functional theory of superconductivity in doped tungsten oxides. *Phys. Rev. Mater.* **3**, 064804 (2019).
46. Marques, M. A. L. et al. Ab initio theory of superconductivity. II. Application to elemental metals. *Phys. Rev. B* **72**, 024546 (2005).
47. Linscheid, A., Sanna, A., Floris, A. & Gross, E. K. U. First-principles calculation of the real-space order parameter and condensation energy density in phonon-mediated superconductors. *Phys. Rev. Lett.* **115**, 097002 (2015).
48. Massidda, S. et al. The role of Coulomb interaction in the superconducting properties of  $\text{CaC}_6$  and H under pressure. *Supercond. Sci. Technol.* **22**, 034006 (2009).
49. Sanna, A. et al. Ab initio Eliashberg theory: making genuine predictions of superconducting features. *J. Phys. Soc. Jpn.* **87**, 041012 (2018).
50. Sano, W., Koretsune, T., Tadano, T., Akashi, R. & Arita, R. Effect of Van Hove singularities on high- $T_c$  superconductivity in  $\text{H}_2\text{S}$ . *Phys. Rev. B* **93**, 094525 (2016).

**Acknowledgements** This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 802533); the Spanish Ministry of Economy and Competitiveness (FIS2016-76617-P); Grant-in-Aid for Scientific Research (number 16H06345, 18K03442 and 19H05825) from the Ministry of Education, Culture, Sports, Science and Technology, Japan; and NCCR MARVEL funded by the Swiss National Science Foundation. Computational resources were provided by the Barcelona Superconducting Center (project FI-2019-1-0031) and the Swiss National Supercomputing Center (CSCS) with project s970.

**Author contributions** The project was conceived by I.E. and J.A.F.-L. The SSCHA was developed by I.E., L.M., R.B., M.C. and F.M. In particular, R.B. developed the method to compute the quantum energy Hessian and the anharmonic phonon dispersions, and L.M. developed the method to perform a quantum relaxation of the lattice parameters. I.E. and F.B. performed the SSCHA calculations. A.S., T.K., T.T., R.A. and J.A.F.-L. conducted studies on structure prediction and superconductivity. All authors contributed to the editing of the manuscript.

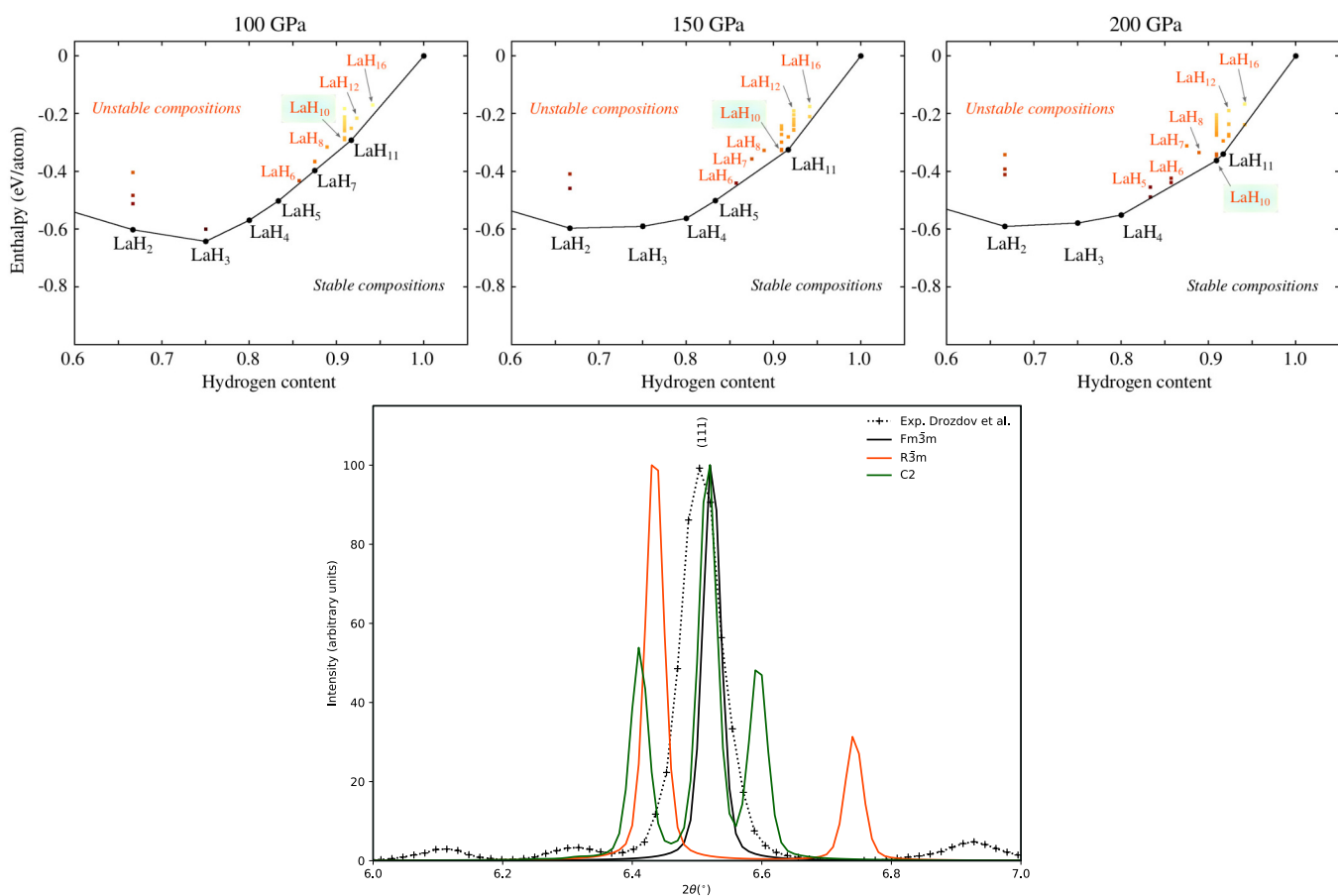
**Competing interests** The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.A.F.-L.

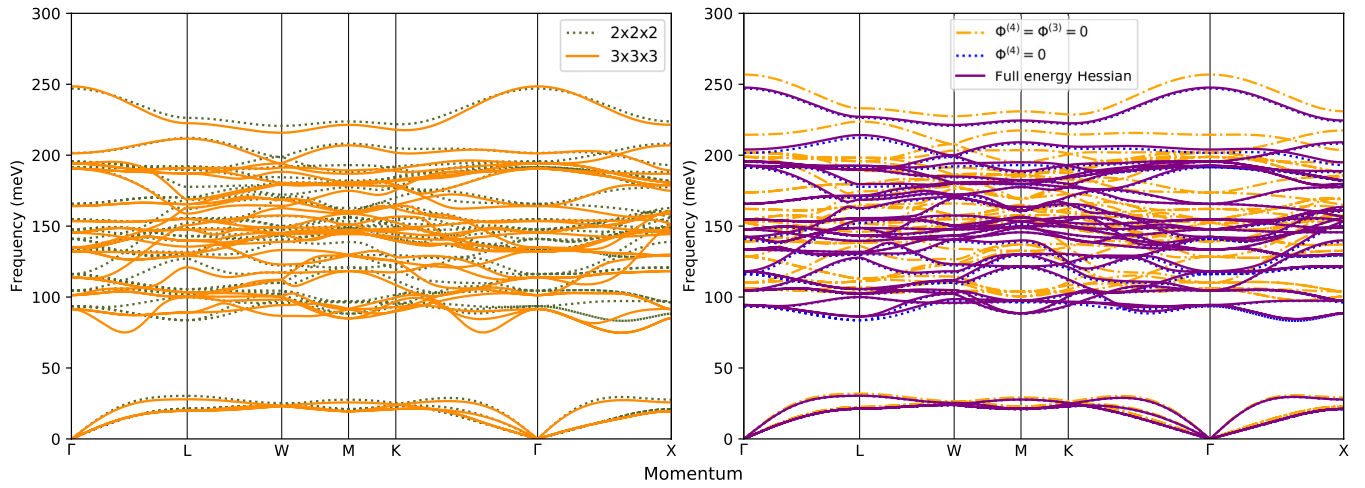
**Peer review information** *Nature* thanks Yanming Ma and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Convex hull of enthalpy formation and diffraction pattern of candidate  $\text{LaH}_{10}$  phases.** Top, classical calculations of enthalpy (without zero-point energy) at different hydrogen contents at 100, 150 and 200 GPa. At low pressure (100 GPa),  $\text{LaH}_{10}$  is not stable and only develops as stable point in the convex hull of enthalpy formation at pressures above about 175 GPa. Bottom, diffraction patterns of different structures at 150 GPa (classical

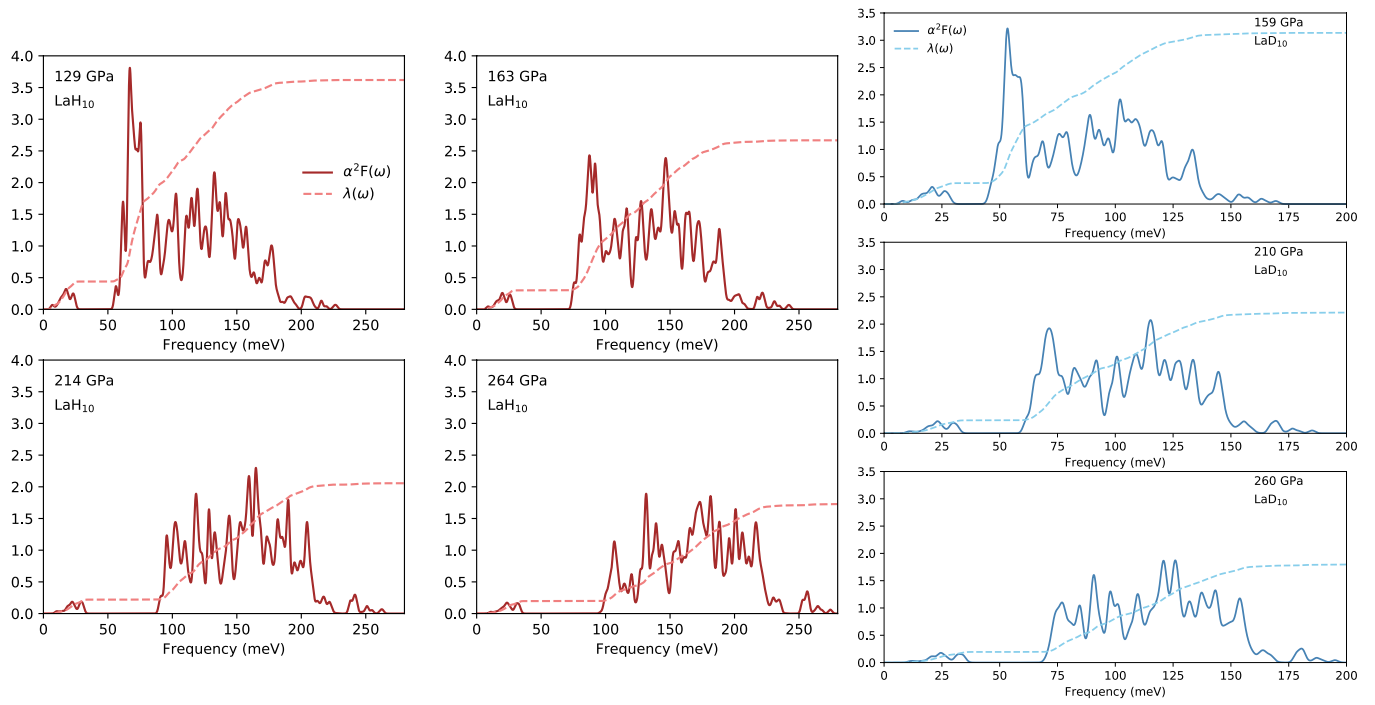
pressure), compared to the experimental data reported in ref. <sup>5</sup> for  $\text{LaH}_{10}$  in the  $\text{Fm}\bar{3}m$  phase at 150 GPa. The pattern is shown in the vicinity of the (111) peak of the  $\text{Fm}\bar{3}m$  phase. This peak is clearly split in the distorted  $\text{C}2$  and  $\text{R}\bar{3}m$  phases predicted classically. The figure provides confirmation that the experimental resolution in ref. <sup>5</sup> would have been sufficient to distinguish between these phases.



**Extended Data Fig. 2 | Convergence of SSCHA-phonon supercells and different anharmonic phonon calculations for  $\text{LaH}_{10}$  at 163 GPa.** Left, the phonon spectra shown are calculated by directly Fourier-interpolating the force constants obtained from the Hessian of  $E(\mathcal{R})$  in a real space  $2 \times 2 \times 2$  and a  $3 \times 3 \times 3$  supercell. The similarity of both phonon spectra obtained by Fourier interpolation indicates that these SSCHA force constants are short-ranged and can be Fourier-interpolated. Right, phonon spectra obtained from the SSCHA energy Hessian of equation (1), making different level of approximations. The purple solid line is the phonon spectrum calculated with the full-energy

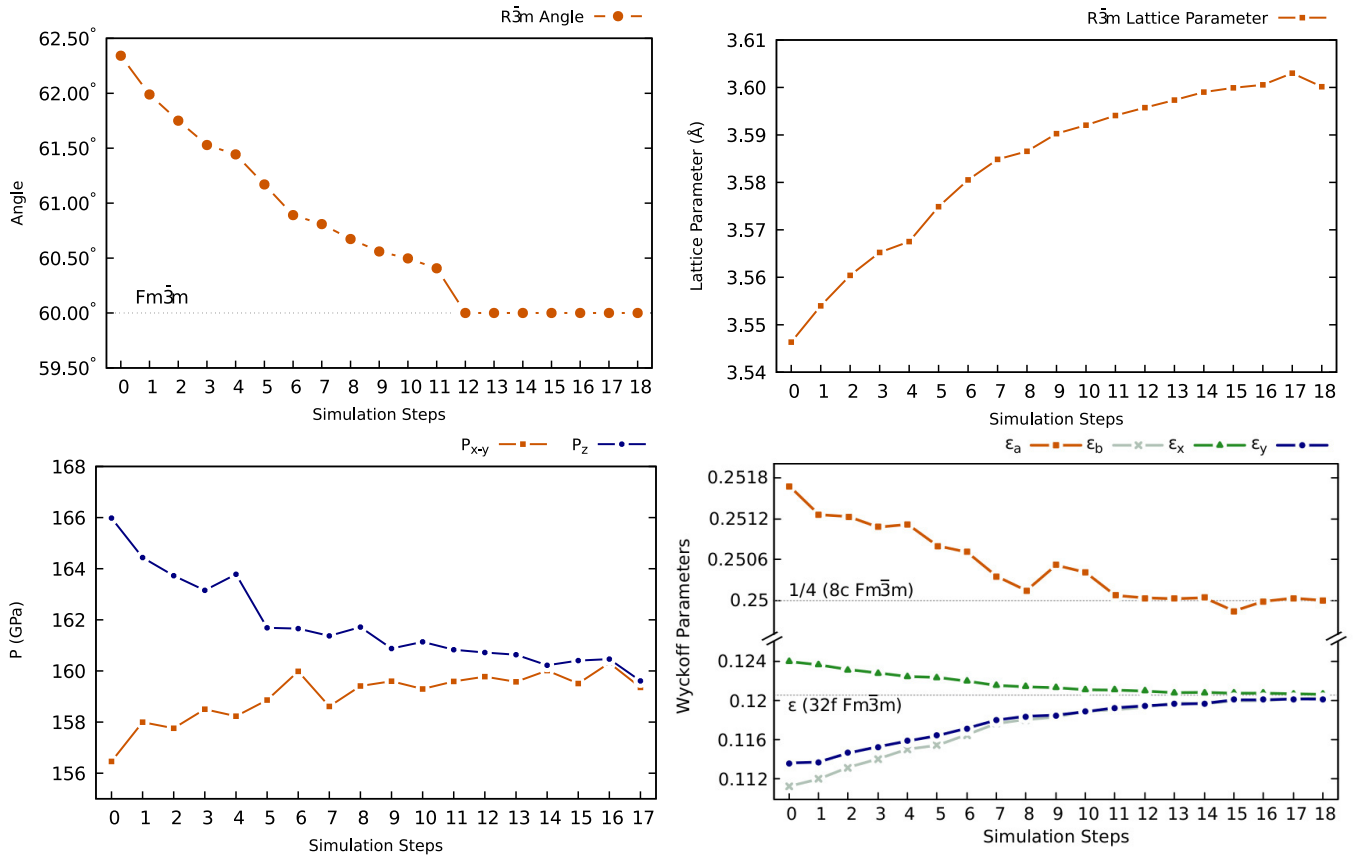
Hessian without any approximation. In the blue dotted spectrum we set  $\Phi^{(4)} = 0$  in the equation. For the orange dash-dotted line we set  $\Phi^{(3)} = \Phi^{(4)} = 0$ , so that the phonon spectra correspond to that arising directly from the SSCHA variational force constants  $\Phi$ . These results clearly show that whereas the effect of  $\Phi^{(3)}$  is important, setting  $\Phi^{(4)} = 0$  has minimal effect. All phonon spectra are obtained by directly Fourier-interpolating the real space anharmonic force constants in a  $2 \times 2 \times 2$  supercell.





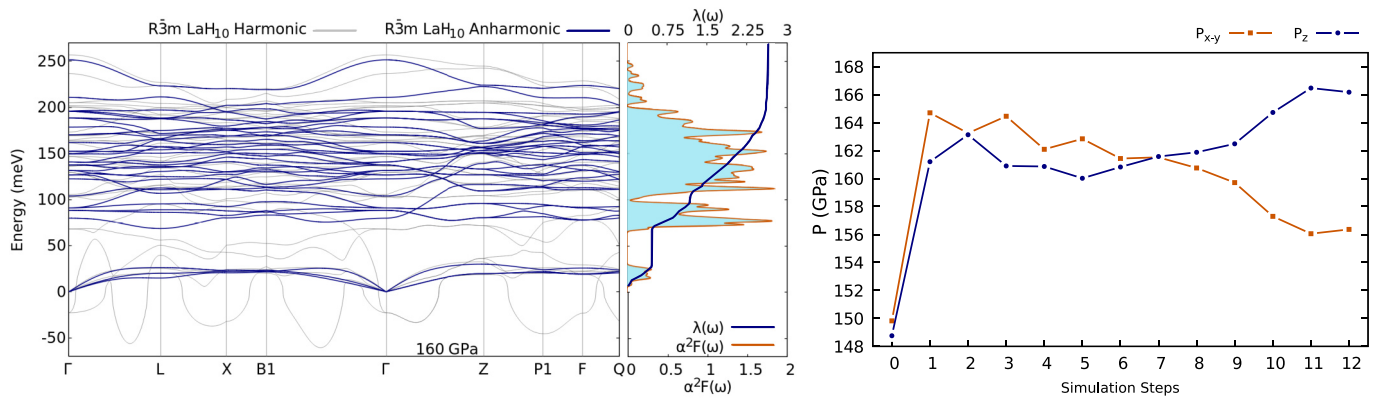
**Extended Data Fig. 3 |  $\alpha^2F(\omega)$  values for the  $Fm\bar{3}m$  phase of  $\text{LaH}_{10}$  and  $\text{LaD}_{10}$ .** Calculated  $\alpha^2F(\omega)$  values for different pressures together with the integrated electron-phonon coupling constant, which is defined as  $\lambda(\omega) = 2 \int_0^\omega d\Omega \alpha^2F(\Omega)/\Omega$ . The results show that high frequency, optical modes of hydrogen are

responsible for the large value of the electron-phonon coupling constant  $\lambda$ . It is worth noting that acoustic modes with La character contribute between 0.2 and 0.5 to  $\lambda$  and cannot be neglected when aiming to estimate an accurate value of  $T_c$ .



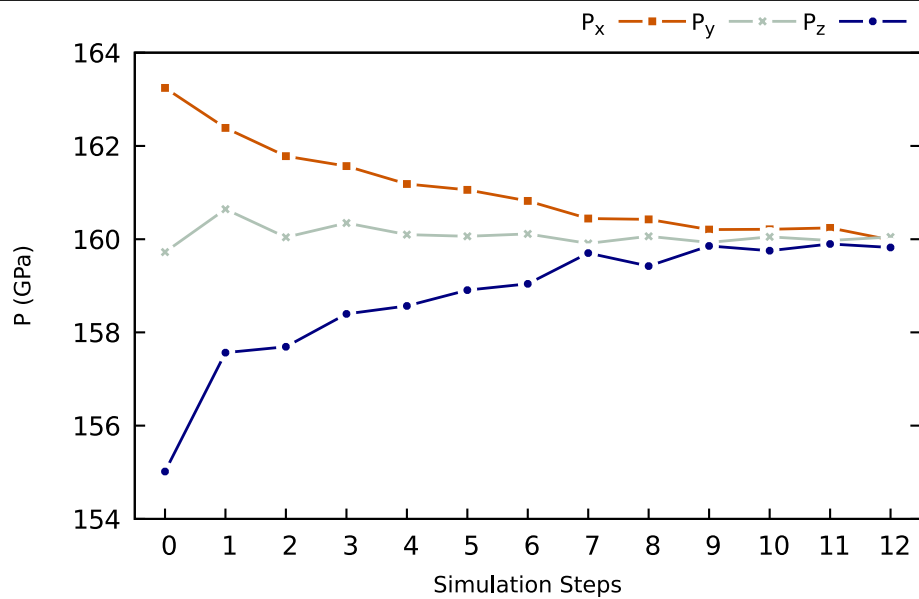
**Extended Data Fig. 4 | Details of the  $R\bar{3}m$   $LaH_{10}$  cell relaxation, including quantum effects.** The initial point for the relaxation is the output from the previous internal relaxation with fixed angle presented in Extended Data Fig. 5. The  $R\bar{3}m$  phase in the rhombohedral description is described by three vectors of the same length ( $a=b=c$ ) and by the angles between them ( $\alpha=\beta=\gamma$ ). The top left panel shows the evolution of the rhombohedral angle and the top right panel shows the evolution of the rhombohedral lattice parameter ( $a=b=c$ ). The progression of the stress tensor in the quantum SSCHA minimization is shown in the bottom left panel. It is clear that at the end of the minimization the structure has an angle of 60°, which matches the angle of an fcc lattice and, in

this case, the stress is isotropic. In the bottom right panel, we show the evolution of the Wyckoff positions in the minimization and we compare it with that of the  $Fm\bar{3}m$  phase. The occupied Wyckoff positions for both  $R\bar{3}m$   $LaH_{10}$  and  $Fm\bar{3}m$   $LaH_{10}$  are summarized in Extended Data Table 2. Here, the evolution of  $\epsilon_a$ ,  $\epsilon_b$ ,  $\epsilon_x$  and  $\epsilon_y$  parameters in the minimization can be seen. The atoms in the first set of 6c positions approach the 8c Wyckoff site of the  $Fm\bar{3}m$  phase, whereas the atoms in the second set of 6c positions and those in 18h sites approach the atoms in the 32f Wyckoff site of the  $Fm\bar{3}m$  phase, where  $\epsilon=0.12053$ .



**Extended Data Fig. 5 | Phonon dispersion in  $R\bar{3}m$ -phase  $\text{LaH}_{10}$  and the anisotropic pressure created in a fixed-cell quantum relaxation.** Left, harmonic and anharmonic phonon spectrum, maintaining a  $62.3^\circ$  rhombohedral angle. The harmonic calculation is performed with the internal atomic positions that yield classical vanishing forces. The anharmonic calculation is performed after relaxing (with the SSCHA) the internal degrees of freedom but maintaining the  $62.3^\circ$  rhombohedral angle. At the harmonic level there are unstable phonon modes even at  $\Gamma$ . Symmetry prevents the relaxation of this structure according to the unstable phonon mode at  $\Gamma$ . The harmonic phonons are calculated at a classic pressure of 150 GPa. Quantum effects add around an extra 10 GPa to the pressure. To the right of the graph is shown the behaviour of  $\lambda(\omega)$  and  $\alpha^2 F(\omega)$  for the anharmonic calculation. Right, pressure along the different Cartesian directions during the SSCHA relaxation of the internal parameters, keeping the rhombohedral angle fixed at  $62.3^\circ$ . At

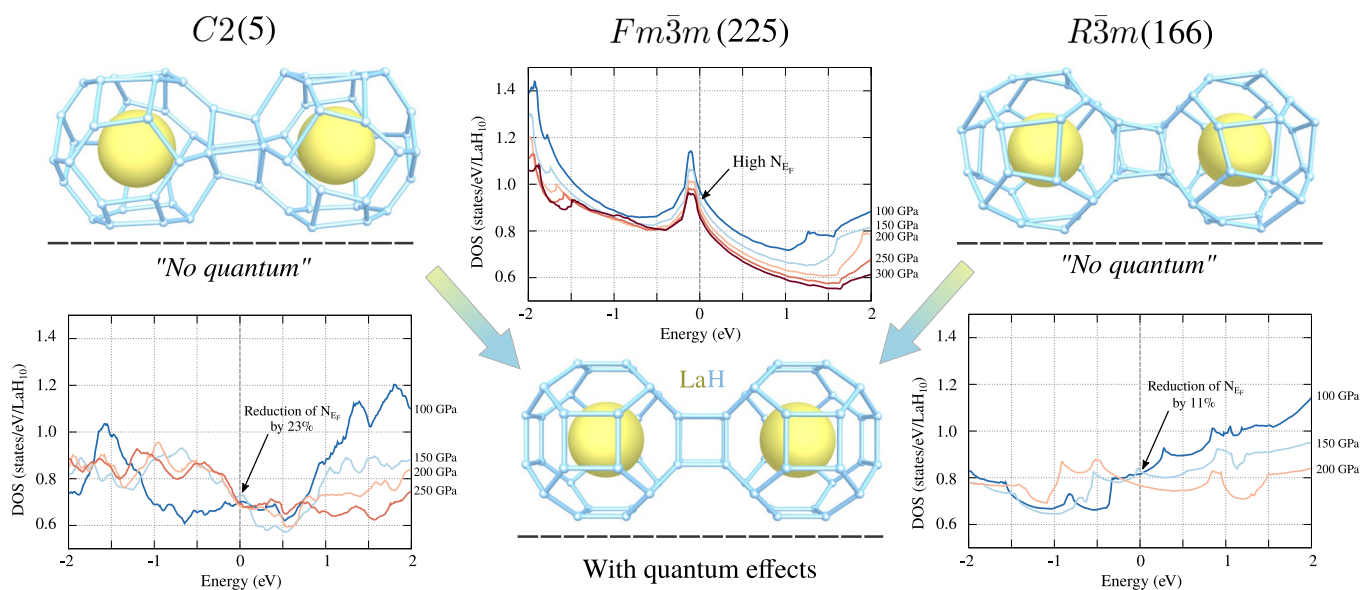
step 0 the pressure reported is obtained directly from  $V(\mathbf{R})$ , neglecting quantum effects. It is isotropic within 1 GPa of difference between the  $x$ - $y$  and  $z$  directions. At each of the other steps it is calculated from the quantum  $E(\mathcal{R})$  and along the minimization it becomes anisotropic. When the minimization stops at step 12—that is, the internal coordinates are at the minimum of the  $E(\mathcal{R})$  for this lattice—the stress anisotropy between the  $z$  and the  $x$ - $y$  directions is about 6%. This clearly indicates that quantum effects act to relax the crystal lattice—in particular, because  $P_z$  is larger—by reducing the rhombohedral angle. It is worth noting that quantum effects increase the total pressure by approximately 10 GPa, which is calculated as  $P = (P_x + P_y + P_z)/3$ . The initial cell parameters before the minimization are  $a = 3.5473398 \text{ \AA}$  and  $\alpha = 62.34158^\circ$ . The initial values of the free Wyckoff parameters, which yield classical vanishing forces and a 150-GPa isotropic stress, are  $\varepsilon_a = 0.26043$ ,  $\varepsilon_b = 0.09950$ ,  $\varepsilon_x = 0.10746$  and  $\varepsilon_y = 0.12810$ . See Extended Data Table 2 for more details.



**Extended Data Fig. 6 | Anisotropic pressure of the C2 phase of  $\text{LaH}_{10}$  in a cell quantum relaxation.** Pressure along the different Cartesian directions is plotted during the SSCHA cell minimization. The target pressure for this minimization is 160 GPa. At the end of the minimization the isotropy of the stress tensor is recovered. A symmetry analysis performed on the structure at

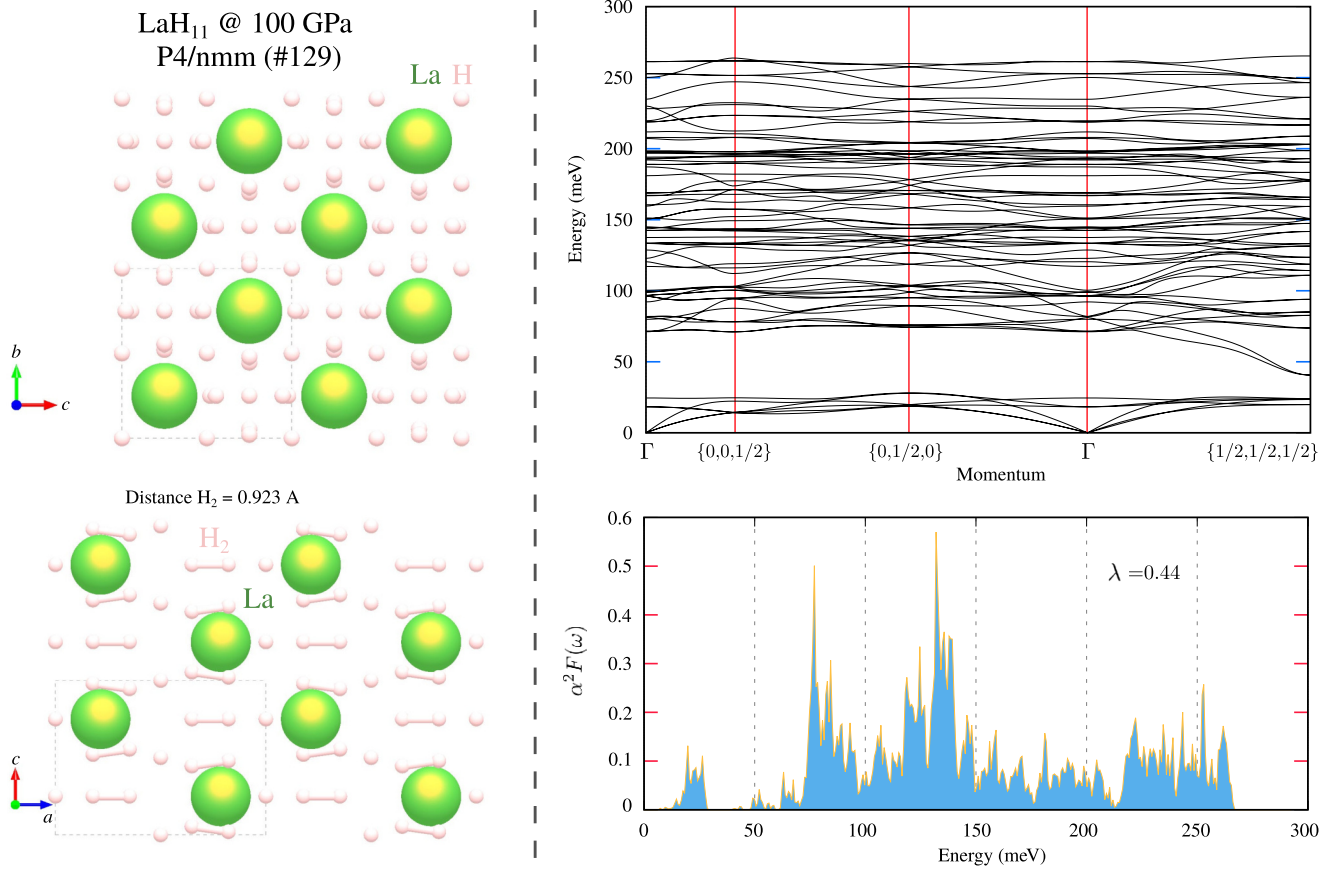
the end of the minimization confirms that the C2  $\text{LaH}_{10}$  evolves in the  $Fm\bar{3}m$ -phase  $\text{LaH}_{10}$ . The initial values  $P_x = 163.2$  GPa,  $P_y = 159.7$  GPa,  $P_z = 155.0$  GPa are obtained by an atomic internal relaxation performed using the SSCHA with a fixed cell.





**Extended Data Fig. 7 | SSCHA minimization on LaH<sub>10</sub> and DOS.** Top left and top right, two initial structures (*C2* and *R3m*) of low enthalpy that were considered in our SSCHA simulations. When considering quantum effects, both structures evolve towards the *Fm3m* structure. The corresponding total electronic DOS at different pressures is plotted for each structure (for comparison, at the same energy scale). The highly symmetric motif (*Fm3m*)

maximizes  $N_{E_F}$ , whereas in distorted structures (*R3m* and *C2*) the occupation at the Fermi level is reduced by more than 23% for *C2* and by 11% for *R3m* (with respect to *Fm3m* at 150 GPa). Values at classical pressures are shown for comparison. Note that the shape of the DOS plot is also strongly modified at different pressures.



**Extended Data Fig. 8 | Details of  $\text{LaH}_{11}$ .** Left, crystal structure of the  $P4/nmm$  phase of  $\text{LaH}_{11}$  at 100 GPa, which is thermodynamically stable in the convex hull. Top right, dispersion of harmonic phonons along the momentum space for  $\text{LaH}_{11}$ ; it is dynamically stable. Bottom right, superconducting Eliashberg

spectrum function ( $\alpha^2 F(\omega)$ ) calculated for  $\text{LaH}_{11}$  at the pressure indicated with harmonic phonons. The  $T_c$  estimated using the Allen–Dynes formula ( $\mu^* = 0.1$ ) is around 7 K at 100 GPa (harmonic phonons).

Extended Data Table 1 | Summary of calculated  $T_c$  values

System	Pressure (GPa)	$\lambda$	$\omega_{log}$ (meV)	$T_{c\mu^*=0.1}^{Mc}$ (K)	$T_{c\mu^*=0.1}^{AD}$ (K)	$T_{c\mu^*=0.1}^{ME}$ (K)	$T_c^{SCDFT}$ (K)
LaH <sub>10</sub>	129	3.62	76.4	171.8	252.6	255.3	230
LaH <sub>10</sub>	163	2.67	96.4	190.4	247.0	242.8	225
LaH <sub>10</sub>	214	2.06	115.5	196.3	235.9	237.9	210
LaH <sub>10</sub>	264	1.73	126.6	189.5	219.2	216.9	201
LaD <sub>10</sub>	159	3.14	63.5	135.0	184.2	180.4	171
LaD <sub>10</sub>	210	2.21	81.7	145.5	176.5	172.9	158
LaD <sub>10</sub>	260	1.80	92.2	142.2	164.6	157.9	151

Values are calculated using different approaches ranging from empirical to fully ab initio: McMillan equation ( $T_{c\mu^*=0.1}^{Mc}$ ), Allen–Dynes-modified McMillan equation ( $T_{c\mu^*=0.1}^{AD}$ ), anisotropic treatment of Migdal–Eliashberg ( $T_{c\mu^*=0.1}^{ME}$ ) and SCDFT ( $T_c^{SCDFT}$ ). Values of  $\lambda$  and  $\omega_{log}$  are also given.

Extended Data Table 2 | Details of the crystal structures

$Fm\bar{3}m$ (C)	$Fm\bar{3}m$ (R)	$R\bar{3}m$ (R)
1 La <b>4b</b> $[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]$	1 La <b>4b</b> $[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]$	1 La <b>3b</b> $[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]$
2 H <b>8c</b> $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$ $[\frac{3}{4}, \frac{3}{4}, \frac{3}{4}]$	2 H <b>8c</b> $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$ $[\frac{3}{4}, \frac{3}{4}, \frac{3}{4}]$	2 H <b>6c</b> $[\epsilon_a, \epsilon_a, \epsilon_a]$ $[-\epsilon_a, -\epsilon_a, -\epsilon_a]$
8 H <b>32f</b> $[\epsilon, \epsilon, \epsilon]$ $[-\epsilon, -\epsilon, -\epsilon]$ $[\epsilon, \epsilon, -\epsilon]$ $[-\epsilon, -\epsilon, \epsilon]$ $[\epsilon, \epsilon, \epsilon]$ $[-\epsilon, -\epsilon, -\epsilon]$ $[\epsilon, \epsilon, -\epsilon]$ $[-\epsilon, -\epsilon, \epsilon]$	8 H <b>32f</b> $[\epsilon, \epsilon, \epsilon]$ $[-\epsilon, -\epsilon, -\epsilon]$ $[-\epsilon, -\epsilon, 3\epsilon]$ $[-\epsilon, 3\epsilon, -\epsilon]$ $[3\epsilon, -\epsilon, -\epsilon]$ $[\epsilon, \epsilon, -3\epsilon]$ $[\epsilon, -3\epsilon, \epsilon]$ $[-3\epsilon, \epsilon, \epsilon]$	2 H <b>6c</b> $[\epsilon_b, \epsilon_b, \epsilon_b]$ $[-\epsilon_b, -\epsilon_b, -\epsilon_b]$ 6 H <b>18h</b> $[-\epsilon_x, -\epsilon_x, 3\epsilon_y]$ $[-\epsilon_y, 3\epsilon_x, -\epsilon_x]$ $[3\epsilon_y, -\epsilon_x, -\epsilon_x]$ $[\epsilon_x, \epsilon_x, -3\epsilon_y]$ $[\epsilon_x, -3\epsilon_y, \epsilon_x]$ $[-3\epsilon_y, \epsilon_x, \epsilon_x]$

Composition (Space group)	Lattice parameters	Wyckoff positions
LaH <sub>10</sub> ( <i>Immm</i> )	$a = 3.58303 \text{ \AA}$ $b = 3.61834 \text{ \AA}$ $c = 5.08749 \text{ \AA}$	La <b>2c</b> [0.50000, 0.50000, 0.00000] H <b>8m</b> [0.75841, 0.00000, 0.11649] H <b>8l</b> [0.00000, 0.75742, 0.87548] H <b>4j</b> [0.50000, 0.00000, 0.74572]
LaH <sub>10</sub> ( <i>C2</i> )	$a = 6.15468 \text{ \AA}$ $b = 3.60628 \text{ \AA}$ $c = 7.23776 \text{ \AA}$ $\beta = 55.71434^\circ$	La <b>4c</b> [0.49244, 0.00070, 0.25292] H <b>4c</b> [0.13978, 0.24567, -0.05243] H <b>4c</b> [0.09798, 0.24122, 0.45027] H <b>4c</b> [0.36015, 0.25590, 0.05238] H <b>4c</b> [0.40204, 0.26021, 0.54971] H <b>4c</b> [-0.09751, 0.00051, -0.05100] H <b>4c</b> [0.86810, 0.00071, 0.43706] H <b>4c</b> [0.88713, 0.00076, 0.69398] H <b>4c</b> [0.87083, 0.00068, 0.19089] H <b>4c</b> [0.73058, 0.00043, 0.88088] H <b>4c</b> [0.76156, 0.00071, 0.36763]
LaH <sub>11</sub> ( <i>P4/nmm</i> )	$a = 3.87435 \text{ \AA}$ $b = 3.87435 \text{ \AA}$ $c = 5.27636 \text{ \AA}$	La <b>2c</b> [0.25000, 0.25000, 0.78577] H <b>4e</b> [0.00000, 0.00000, 0.50000] H <b>8i</b> [0.25000, -0.02052, 0.17824] H <b>8i</b> [0.25000, 0.55418, 0.35160] H <b>2a</b> [0.75000, 0.25000, 0.00000]

Top, the table summarizes the occupied Wyckoff positions for different structures found with the minima hopping method and used in SSCHA for minimization. We describe the Wyckoff positions using crystal coordinates, so that the  $[x, y, z]$  coordinate should be understood as an  $x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$  atomic position with  $\mathbf{a}, \mathbf{b}, \mathbf{c}$  the lattice vectors. For the  $R\bar{3}m$  phase we use the rhombohedral lattice (R), where the three lattice vectors have the same length ( $\mathbf{a} = \mathbf{b} = \mathbf{c}$ ) and the angle between them is the same ( $\alpha = \beta = \gamma$ ). The  $Fm\bar{3}m$  phase is described both in this rhombohedral description (R) and, for comparison, in the standard cubic conventional lattice (C). In the  $Fm\bar{3}m$  phase the lanthanum atom is described by the  $4b$  sites, two hydrogen atoms occupy the  $8c$  sites, and the remaining eight hydrogen atoms occupy the  $32f$  sites. Most of the atomic positions are fixed by symmetry, and overall the  $Fm\bar{3}m$  structure can be described by one single free parameter ( $\epsilon$ ). In the  $R\bar{3}m$  phase the lanthanum atom is locked in the  $3b$  sites, two pairs of hydrogen atoms occupy the  $6c$  sites and the remaining six hydrogen atoms occupy the  $18h$  sites. In this case symmetry allows for more freedom and overall the structure of the  $R\bar{3}m$  phase can be described by four free parameters ( $\epsilon_a, \epsilon_b, \epsilon_x$  and  $\epsilon_y$ ). The bottom table shows lattice parameters and atomic coordinates for LaH<sub>10</sub> (*Immm*) and LaH<sub>10</sub> (*C2*) at 150 GPa and LaH<sub>11</sub> *P4/nmm* at 100 GPa. These pressures are estimated classically. The positions below give vanishing forces at classical level.



# Spin current from sub-terahertz-generated antiferromagnetic magnons

<https://doi.org/10.1038/s41586-020-1950-4>

Received: 11 January 2019

Accepted: 22 October 2019

Published online: 27 January 2020

Junxue Li<sup>1</sup>, C. Blake Wilson<sup>2,3</sup>, Ran Cheng<sup>1,4</sup>, Mark Lohmann<sup>1</sup>, Marzieh Kavand<sup>2,3</sup>, Wei Yuan<sup>1</sup>, Mohammed Aldosary<sup>1</sup>, Nikolay Agladze<sup>2,3</sup>, Peng Wei<sup>1</sup>, Mark S. Sherwin<sup>2,3</sup> & Jing Shi<sup>1\*</sup>

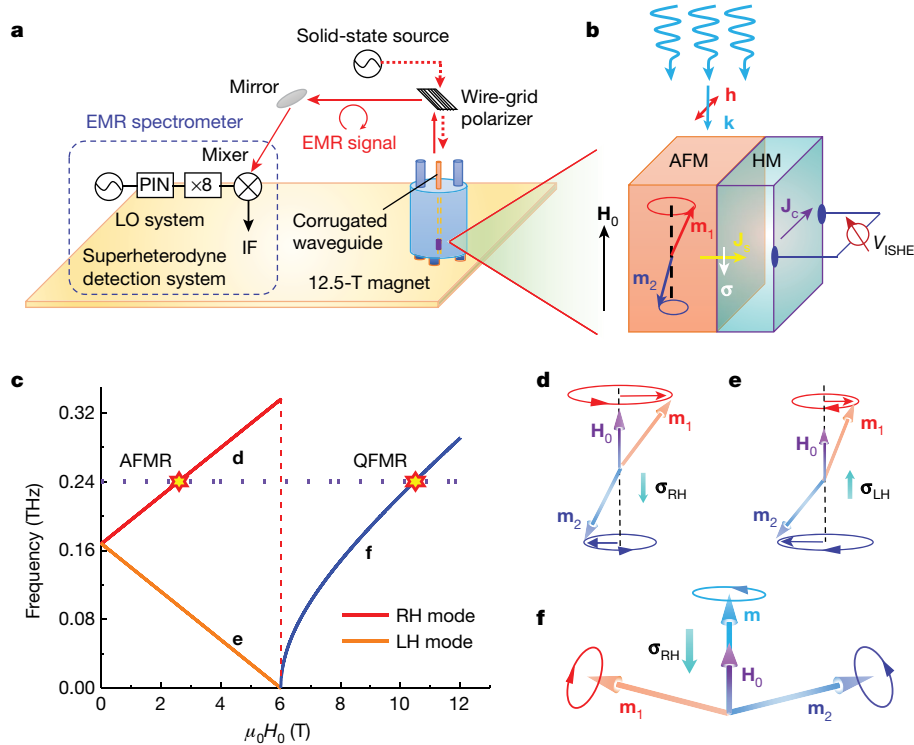
Spin dynamics in antiferromagnets has much shorter timescales than in ferromagnets, offering attractive properties for potential applications in ultrafast devices<sup>1–3</sup>. However, spin-current generation via antiferromagnetic resonance and simultaneous electrical detection by the inverse spin Hall effect in heavy metals have not yet been explicitly demonstrated<sup>4–6</sup>. Here we report sub-terahertz spin pumping in heterostructures of a uniaxial antiferromagnetic Cr<sub>2</sub>O<sub>3</sub> crystal and a heavy metal (Pt or Ta in its  $\beta$  phase). At 0.240 terahertz, the antiferromagnetic resonance in Cr<sub>2</sub>O<sub>3</sub> occurs at about 2.7 tesla, which excites only right-handed magnons. In the spin-canting state, another resonance occurs at 10.5 tesla from the precession of induced magnetic moments. Both resonances generate pure spin currents in the heterostructures, which are detected by the heavy metal as peaks or dips in the open-circuit voltage. The pure-spin-current nature of the electrically detected signals is unambiguously confirmed by the reversal of the voltage polarity observed under two conditions: when switching the detector metal from Pt to Ta, reversing the sign of the spin Hall angle<sup>7–9</sup>, and when flipping the magnetic-field direction, reversing the magnon chirality<sup>4,5</sup>. The temperature dependence of the electrical signals at both resonances suggests that the spin current contains both coherent and incoherent magnon contributions, which is further confirmed by measurements of the spin Seebeck effect and is well described by a phenomenological theory. These findings reveal the unique characteristics of magnon excitations in antiferromagnets and their distinctive roles in spin–charge conversion in the high-frequency regime.

Owing to their terahertz spin dynamics and absence of net magnetization, antiferromagnetic (AFM) materials offer unique advantages for ultrafast and robust spin-based nanoscale device applications<sup>10–16</sup>. A prerequisite for practical AFM-based spintronics is the generation and electrical detection of pure spin currents. Cheng et al.<sup>4</sup> and Johansen et al.<sup>5</sup> proposed to generate coherent magnon spin currents by inducing uniform spin precession at the antiferromagnetic resonance (AFMR) with terahertz radiation. This mechanism works in the collinear AFM phase at arbitrary magnetic fields (even zero field) below the spin-flop transition. Ross et al.<sup>6</sup> detected d.c. voltages at the AFMR in MnF<sub>2</sub>, but concluded that the main d.c. voltage was from the microwave rectification effect or heating-related thermoelectric electromotive force (EMF). Clear AFM spin pumping has yet to be experimentally established. Here we demonstrate the generation and simultaneous electrical detection of pure spin currents in a uniaxial AFM material, Cr<sub>2</sub>O<sub>3</sub>. The former is accomplished by driving the AFM spin precession into resonance using linearly polarized sub-terahertz radiation. Through the inverse spin Hall effect (ISHE), the resonantly generated spin current is converted into a d.c. voltage.

Cr<sub>2</sub>O<sub>3</sub> is a uniaxial AFM insulator with the easy axis along the *c* axis of the hexagonal lattice<sup>17</sup> (Extended Data Fig. 1). Because of its relatively

simple spin structure and accessible AFMR frequency (about 0.165 THz at 0 K) and spin-flop field (6.0 T at 0 K)<sup>18,19</sup>, Cr<sub>2</sub>O<sub>3</sub> is chosen for this study. As shown in Fig. 1a, 0.240-THz continuous microwaves are generated by a solid-state source and propagate into a corrugated waveguide at the centre of a 12.5-T superconducting magnet. The Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) slab is mounted on a piece of sapphire secured on a Teflon stage located at the exit of the waveguide. Figure 1b depicts the sample structure and measurement geometry. The sample is oriented with the *c* axis parallel to both the d.c. magnetic field, **H**<sub>0</sub>, and the microwave propagation direction, **k**. The microwave magnetic-field component **h** is kept in the sample plane and perpendicular to the *c* axis of Cr<sub>2</sub>O<sub>3</sub>. Similarly to conventional ferromagnetic spin pumping, the AFM spins in Cr<sub>2</sub>O<sub>3</sub> are driven into resonance at a fixed, but much higher, microwave frequency by sweeping the d.c. magnetic field. A pure spin current is then injected into the adjacent heavy-metal layer, which in turn produces a charge current in Pt, Ta or the Pt–Ta hybrid channel (Methods) through the ISHE and results in an open-circuit d.c. voltage, *V*<sub>ISHE</sub>. Microwaves reflected by the sample are detected by a superheterodyne receiver and recorded as electron magnetic resonance (EMR) signals (Methods)<sup>20,21</sup>. Both EMR and electrical voltage signals are simultaneously measured using the standard lock-in technique.

<sup>1</sup>Department of Physics and Astronomy, University of California, Riverside, CA, USA. <sup>2</sup>Physics Department, University of California, Santa Barbara, CA, USA. <sup>3</sup>Institute for Terahertz Science and Technology, University of California, Santa Barbara, CA, USA. <sup>4</sup>Department of Electrical and Computer Engineering, University of California, Riverside, CA, USA. \*e-mail: jing.shi@ucr.edu



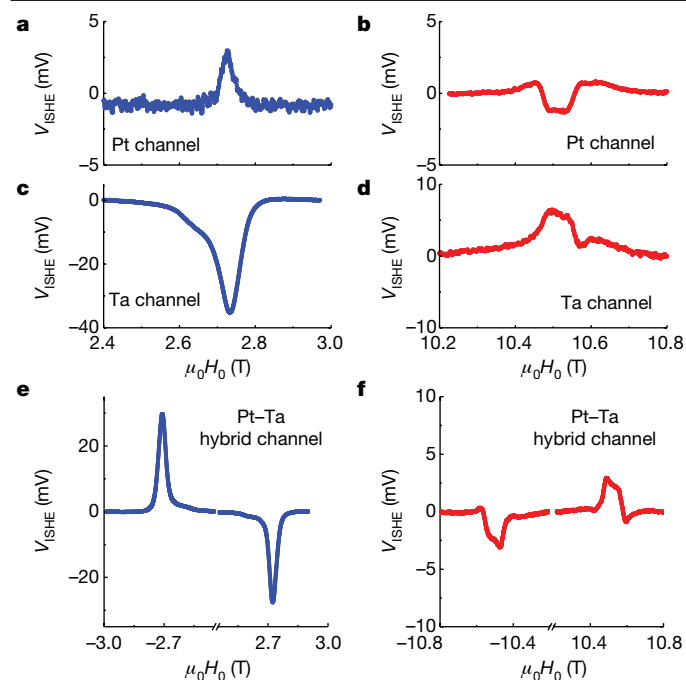
**Fig. 1 | AFM spin dynamics and pure spin current in an AFM/heavy-metal heterostructure.** **a**, Continuous-wave (CW) EMR system (see Methods). A 0.240-THz CW is generated by a solid-state source and polarized by a wire-grid polarizer before entering the waveguide. The sample is loaded into a continuous-flow cryostat mounted in the room-temperature bore of the magnet. CW EMR measurements are carried out using source-intensity modulation with a frequency of 13.037 Hz. The reflected CWs are measured by a superheterodyne detection system using a local oscillator (LO) and a Schottky subharmonic mixer to mix the 0.240-THz signal to 10 GHz. Then, a home-built intermediate-frequency (IF) stage amplifies and mixes this signal

down to baseband. The resulting signal is measured in quadrature with a pair of lock-in amplifiers. PIN, p-i-n switch. **b**, Sample structure and spin-current injection. Two sublattice magnetic moments,  $\mathbf{m}_1$  and  $\mathbf{m}_2$ , are excited into resonance, and the spin current  $\mathbf{J}_s$  with spin polarization  $\sigma$  generates the charge current  $\mathbf{J}_c$  in the heavy-metal (HM) layer. **c**, Magnetic-resonance frequency as a function of the magnetic field  $\mu_0 H_0$  ( $\mu_0$ , magnetic permeability constant) applied along the  $c$  axis of  $\text{Cr}_2\text{O}_3$  at 0 K. The labels **d**, **e** and **f** correspond to the panels at right. **d–f**, The eigenmodes of the RH AFMR (**d**), the LH AFMR (**e**), and the QFMR (**f**), as labelled in **c**.  $\sigma_{\text{RH}}$  and  $\sigma_{\text{LH}}$  are the spin polarizations associated with RH and LH chirality, respectively.

Figure 1c summarizes the magnetic resonance frequency as a function of  $\mathbf{H}_0$  when the magnetic field is applied along the  $c$  axis of  $\text{Cr}_2\text{O}_3$ . Below the spin-flop field of 6.0 T, there are two distinct branches corresponding to the two eigenmodes of the AFM spin-wave excitations or magnons, namely, right-hand (RH) and left-hand (LH) spin precessions with opposite chiralities<sup>14,18,19</sup>. At  $\mathbf{H}_0 = 0$ , these two modes are degenerate at  $\omega_m/(2\pi) = 0.165$  THz, as shown in Fig. 1d, e ( $\omega_m$  is the magnon angular frequency at zero magnetic field). This is fundamentally different from ferromagnetic materials, in which the sole magnon mode is RH. Because these two modes carry equal but opposite angular momenta,  $\pm\hbar$  ( $\hbar$ , reduced Planck constant; ref.<sup>14</sup>), the net spin angular momentum is zero if they are equally populated. When  $\mathbf{H}_0$  is applied along the  $c$  axis of  $\text{Cr}_2\text{O}_3$ , the degeneracy between RH and LH modes is lifted, and the frequencies of the two branches are given by<sup>19</sup>  $\omega/\gamma = \sqrt{2H_E H_A + (H_0 \alpha/2)^2} \pm H_0(1 - \alpha/2)$ , where  $\gamma = 28 \text{ GHz T}^{-1}$  is the gyromagnetic ratio,  $H_E$  and  $H_A$  are effective fields of the inter-sublattice exchange interaction and the easy-axis anisotropy,  $\alpha = \chi_{\parallel}/\chi_{\perp}$  is the ratio of the magnetic susceptibilities in the parallel and perpendicular directions, and the + (–) sign refers to the RH (LH) mode. At low temperatures,  $\alpha \approx 0$  and  $\omega/\gamma = \sqrt{2H_E H_A} \pm H_0$ , which is represented by the two straight lines in Fig. 1c. In our experiments, the microwaves are linearly polarized, and the frequency is held at 0.240 THz (horizontal dashed line) while  $\mathbf{H}_0$  is swept; therefore, only the RH mode can be excited when the upper branch intercepts the horizontal dashed line. For  $\text{Cr}_2\text{O}_3$ , the low-temperature resonance field of the RH mode is estimated to be about 2.7 T, which is confirmed by our EMR experiment (Supplementary Information Note I).

As the  $\mathbf{H}_0$  strength reaches the spin-flop field, spins in both sublattices switch abruptly to align nearly perpendicular to  $\mathbf{H}_0$  with a small inclination<sup>22</sup>. We observe an EMR feature at the spin-flop transition (Supplementary Information Note I). A new resonance mode emerges above the spin-flop field, as depicted in Fig. 1f. In this mode, the total magnetic moment  $\mathbf{m} = \mathbf{m}_1 + \mathbf{m}_2$  precesses around  $\mathbf{H}_0$  with RH chirality<sup>19</sup>. This is essentially equivalent to the ferromagnetic resonance mode in ferromagnets. We call it quasi-ferromagnetic resonance (QFMR) mode to distinguish it from the ferromagnetic resonance mode of the fully spin-aligned state, which could only be accessed at extremely high magnetic fields. The QFMR frequency is given by<sup>19</sup>  $\omega/\gamma = \sqrt{H_0^2 - 2H_E H_A}$ . For  $\omega/(2\pi) = 0.240$  THz, the QFMR is estimated to occur at ~10.5 T in  $\text{Cr}_2\text{O}_3$  (see Supplementary Information Note I for the EMR signal at the QFMR). By sweeping  $\mathbf{H}_0$  up to 12 T, we can excite both the RH AFMR and QFMR modes in  $\text{Cr}_2\text{O}_3$ .

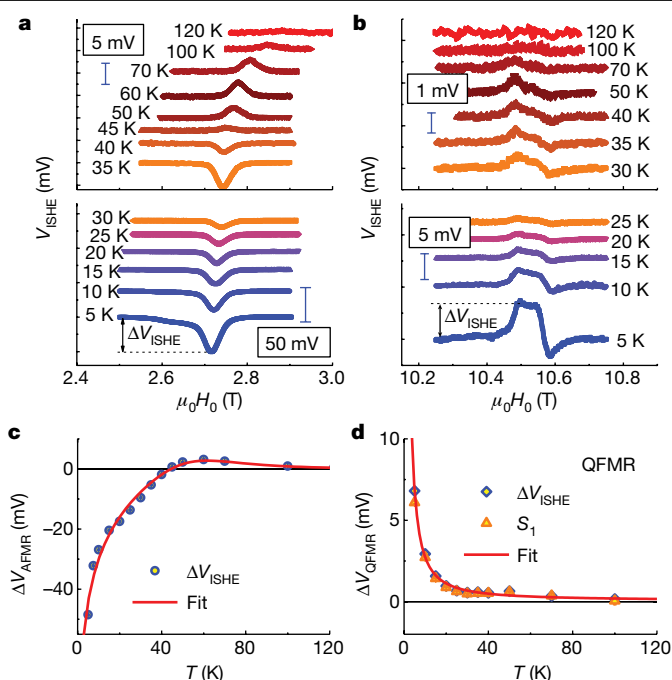
At these resonances, uniformly precessing spins in  $\text{Cr}_2\text{O}_3$  form a  $\mathbf{k} = 0$  magnon reservoir. Similar to ferromagnetic spin pumping, when the magnon reservoir is in contact with a heavy-metal layer such as Pt, a pure spin current flows across the interface via magnon–electron interactions and is consequently converted to an open-circuit d.c. voltage in Pt due to the ISHE. We observed electrical voltage signals at both resonance fields identified by EMR signals. To distinguish the pure-spin-current effect from other spurious effects, we use Pt and Ta as two independent detection channels (Extended Data Fig. 2). Because of the opposite sign in their spin Hall angles<sup>7–9</sup>, the same pure spin current must produce voltage signals with opposite polarities in the Pt and Ta



**Fig. 2 | ISHE signals at the AFMR and the QFMR.** **a, c,** Electrical voltages from the Pt channel (**a**) and the Ta channel (**c**) at the AFMR at 5 K. **b, d,** Voltages from the Pt channel (**b**) and the Ta channel (**d**) at the QFMR at 5 K. **e, f,** Voltages from the Pt-Ta hybrid channel at the AFMR (**e**) and the QFMR (**f**) for both positive and negative fields at 10 K. The ISHE voltage is dominated by the Ta channel at low temperatures.

channels. On the other hand, any spurious electrical signal generated by microwave rectification or heating-related thermoelectric EMF may maintain the same polarity in Pt and Ta. Indeed, sharp voltage features are unmistakably resolved at 2.7 T in two independent Pt and Ta channels, as shown in Fig. 2a, c. The resonance field corresponds well to the AFMR field identified by EMR. More importantly, the Pt and Ta channels register opposite AFMR voltages at 5 K. At 10.5 T, where the QFMR occurs, the same opposite polarity is observed, as shown in Fig. 2b, d. The opposite voltage polarity between the Pt and Ta channels at both the AFMR and the QFMR is a defining characteristic of the resonantly generated pure spin current. Because the ISHE voltage is proportional to the resistivity of the detecting channel, providing the same spin-charge conversion efficiency, and Ta is much more resistive (by more than a factor of 10) than Pt, we expect a larger voltage response in Ta (Supplementary Information Note II).

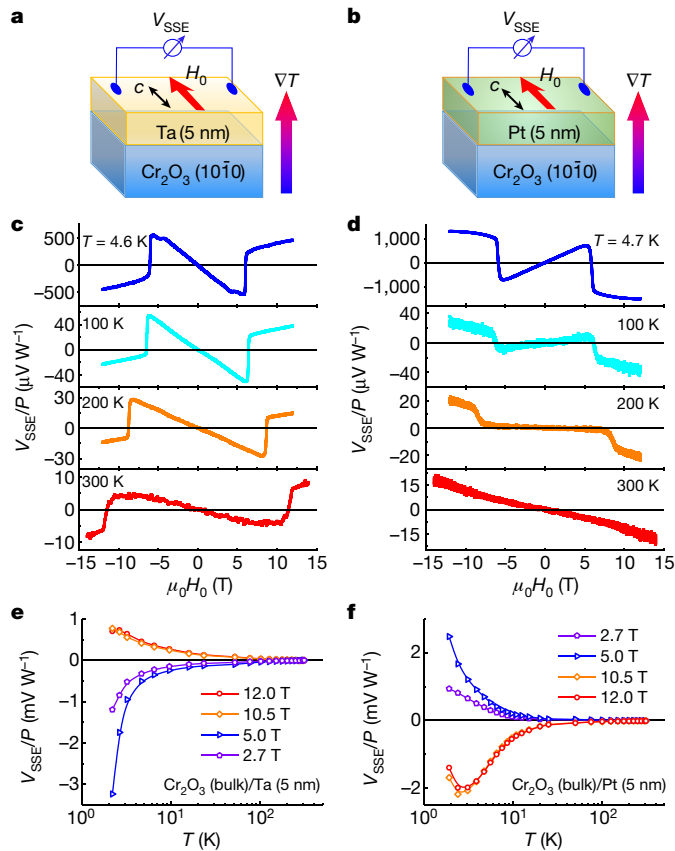
After confirming the opposite voltage polarities in the independent Ta-only and Pt-only channels, we wire-bond the neighbouring Pt and Ta strips in series (Extended Data Fig. 2c) to form a single long hybrid channel. All Pt and Ta strips on the chip are connected to further enhance the total voltage signal output. In the following experiments, unless otherwise specified, we adopt this hybrid detector geometry for electrical detection. To corroborate the pure-spin-current nature of the d.c. voltage signals, we reverse the direction of  $\mathbf{H}_0$ , and therefore the spin polarization of the pumped spin current. We observe a completely inverted voltage signal on the negative-field side (Fig. 2e). The same complete voltage inversion is observed for the QFMR despite the complicated line shape (shown in Fig. 2f). We note that heating-related thermoelectric EMF cannot produce voltage sign reversal when the magnetic field is reversed. Hence, the observation of the two sign reversals leads us to unambiguously conclude that the d.c. voltage in heavy-metal detectors stems exclusively from pure spin current generated by resonant magnon excitations. Furthermore, we confirm that the ISHE voltages have a linear dependence on the microwave power (Extended Data Fig. 3).



**Fig. 3 | Temperature dependence of ISHE signals.** **a,** ISHE voltage signal at the AFMR. **b,** ISHE voltage signal at the QFMR. **c,** Temperature dependence of the ISHE peak height  $\Delta V_{\text{ISHE}}$  (as indicated in **a**) at the AFMR (symbols) and the best fit (line) using equation (S16) in Supplementary Information. **d,** Temperature dependence of the ISHE peak height  $\Delta V_{\text{ISHE}}$  (as indicated in **b**) and of  $S_1$  (symbols): the magnitude of the main symmetric Lorentzian function (see Supplementary Information section III). The red solid line is the best fit using equation (S17) in Supplementary Information. The error bars represent the range of the off-resonance voltage signals. Error bars in **c** and **d** are smaller than the symbols.

A more rigorous examination of the low-temperature behaviours of the AFMR and QFMR shown in Fig. 2 reveals something counterintuitive. With the same detecting heavy metal, the ISHE voltage feature at the AFMR and the main feature at the QFMR have opposite signs, which contradicts the simple picture of coherent spin pumping. As schematically illustrated in Fig. 1d, f, the RH AFMR and the QFMR eigenmodes exhibit the same spin polarization; hence, the resulting ISHE voltages should also have the same polarity for a fixed heavy metal. The apparent contradiction implies that coherent spin pumping is not the sole mechanism. To better understand the origin of the sign discrepancy, we perform ISHE voltage measurements over a wide range of temperatures. Figure 3a shows a plot of the ISHE voltage at the AFMR for positive magnetic fields from 5 K to 120 K. The voltage signal is negative at low temperatures and becomes smaller as the temperature is raised. It crosses zero at about 45 K and stays positive at higher temperatures, until it finally diminishes at around 120 K. A similar sign change is also observed for negative magnetic fields (Extended Data Fig. 4). By contrast, the main ISHE peak at the QFMR (Fig. 3b) always stays positive, and decreases monotonically with increasing temperature until it finally vanishes above 100 K. We also confirmed the sign-change pattern of the ISHE signals by comparing the Pt-only and Ta-only channels (Extended Data Fig. 5).

After carefully analysing the line shape of the d.c. voltage signals at the QFMR at all temperatures (Supplementary Information Note III), we plotted the magnitudes of both the AFMR voltage and the main QFMR peak voltage (Fig. 3c, d). The contrast ends at low temperatures. Clearly, the ISHE voltage at both resonances disappears far below the Néel temperature of the  $\text{Cr}_2\text{O}_3$  crystal (307 K)<sup>18,19</sup>. We note that the EMR signal at the AFMR is still observable up to 288 K (see



**Fig. 4 | SSE from incoherent AFM magnons.** **a, b**, SSE measurements in Cr<sub>2</sub>O<sub>3</sub>(1010)/Ta (**a**) and Cr<sub>2</sub>O<sub>3</sub>(1010)/Pt (**b**) heterostructures. The vertical temperature gradient  $\nabla T$  is generated by an on-chip heater. The magnetic field  $H_0$  is applied in plane and along the  $c$  axis. **c, d**, Field dependence of the SSE signal, normalized to the heating power, at 4.6 K, 100 K, 200 K and 300 K in Cr<sub>2</sub>O<sub>3</sub>/Ta (**c**) and Cr<sub>2</sub>O<sub>3</sub>/Pt (**d**) heterostructures. **e, f**, Temperature dependence of the SSE signal, normalized to the heating power, under magnetic fields of 2.7 T, 5.0 T, 10.5 T and 12 T in Cr<sub>2</sub>O<sub>3</sub>/Ta (**e**) and Cr<sub>2</sub>O<sub>3</sub>/Pt (**f**) heterostructures.

Supplementary Fig. 4), indicating that microwave absorption remains active at least up to room temperature. Because the ISHE voltage depends on the efficiency of the spin–charge conversion, we believe that the quality of the interface may be responsible for the disappearance of the ISHE signals at a lower temperature than the ordering temperature of Cr<sub>2</sub>O<sub>3</sub>.

The stark contrast between the temperature dependence of the AFMR- and QFMR-induced spin currents indicates that coherent spin pumping alone is inadequate to explain these behaviours. The reasons include: (1) no sign change of the dependence of the spin Hall angle of Pt or Ta on the temperature has ever been reported<sup>7,8</sup>; (2) the absence of sign change for the QFMR further confirms point (1); (3) the spin-polarization direction of the resonant mode is fixed and cannot change with temperature. We propose the following mechanism to explain this unusual temperature dependence. As illustrated in Fig. 1c, whereas the coherently driven AFMR selectively excites the RH mode, thermal excitations of incoherent magnons prefer the LH mode, which has a lower energy. As a result, thermally driven LH magnons compete with coherent RH magnons. When the contribution of the former exceeds that of the latter, it causes a sign change of the total ISHE voltage. Therefore, the sign change strongly suggests rapid thermalization of coherent magnons into incoherent magnons. This process results in an increased effective magnon temperature, which in turn raises the lattice temperature via magnon–phonon scattering. In fact, we observed a temperature rise at the AFMR by monitoring the resistance of the Pt–Ta hybrid channel, which serves as a sensitive

thermometer (Supplementary Information Note IV). A similar mechanism has recently been proposed for ferromagnetic spin pumping, which can result in resonance-induced heating<sup>23,24</sup>.

The increased incoherent magnon population at the resonances must be accompanied by an additional spin current flowing across the metal–Cr<sub>2</sub>O<sub>3</sub> interface, and consequently an additional ISHE voltage of opposite polarity. This process is essentially the same as the spin Seebeck effect (SSE), in which the incoherent magnon diffusion is driven by a temperature gradient. To better understand the incoherent magnon contribution, we performed independent SSE measurements in the same Cr<sub>2</sub>O<sub>3</sub>/Ta and Cr<sub>2</sub>O<sub>3</sub>/Pt heterostructures, where coherent magnons were completely eliminated. As illustrated in Fig. 4a, b, we measured the SSE-induced ISHE voltages by sweeping the magnetic field along the  $c$  axis of the Cr<sub>2</sub>O<sub>3</sub> crystal under a vertical temperature gradient generated by a heater on top of the crystal. Figure 4c, d displays the field dependence of the SSE signals in Cr<sub>2</sub>O<sub>3</sub>/Ta and Cr<sub>2</sub>O<sub>3</sub>/Pt. First, there is a large SSE signal at low fields. At 6 T—that is, the spin-flop field—the SSE voltage undergoes an abrupt jump and switches sign. The sign switch is exactly what is expected from the magnon energy diagram shown in Fig. 1c. Below the spin-flop transition, the LH-magnon branch has a lower energy, and thus the total thermal magnon population is dominated by LH magnons at low temperatures. Above the spin-flop transition, however, the canted state supports only RH magnons; therefore, the SSE must change sign across the spin-flop transition. The same behaviour was also observed in heterostructures containing epitaxial Cr<sub>2</sub>O<sub>3</sub> thin films (Extended Data Fig. 6). The SSE signal in Cr<sub>2</sub>O<sub>3</sub>/Pt is simply inverted compared to that in Cr<sub>2</sub>O<sub>3</sub>/Ta for the same reason as in the resonance data. Our SSE signal is quite different from what was reported by Seki et al.<sup>25</sup>, which might be caused by the different Cr<sub>2</sub>O<sub>3</sub>/heavy-metal interface properties (as demonstrated in Extended Data Fig. 7). The temperature dependences of both samples are summarized in Fig. 4e, f for magnetic fields below and above the spin-flop transition and temperatures of up to 310 K. It is interesting to note that the incoherent magnon contribution to the ISHE voltage in Cr<sub>2</sub>O<sub>3</sub> alone does not cause any sign change, which is different from the behaviours of some ferrimagnets<sup>26,27</sup> (see discussion in Supplementary Information Note V). As the temperature is decreased, the magnitude of the SSE voltage below the spin-flop field increases precipitously. This trend is the same as that of the AFMR ISHE voltage. Both can be explained by a decreased equilibrium population of RH magnons. At the AFMR, as the temperature is lowered, the increased LH-magnon contribution balances out that of the coherently excited RH magnons at ~45 K. In fact, as the temperature is decreased further, the net LH-magnon population should reach the maximum, and an SSE voltage peak emerges<sup>28</sup>. The peak occurs approximately at temperature  $T_p$  when  $k_B T_p \approx \hbar \omega_m$ , where  $k_B$  is the Boltzmann constant. This overall SSE characteristic has been observed in other uniaxial AFM materials, such as MnF<sub>2</sub> and FeF<sub>2</sub><sup>29,30</sup>. Compared to these two materials, Cr<sub>2</sub>O<sub>3</sub> has a lower  $\omega_m$ ; hence, the SSE peak should occur at an even lower  $T_p$ . We note that the SSE peak was not captured in previous experiments because  $T_p$  was outside their temperature range<sup>25</sup>, but it appears at ~2.3 K in our experiment, as shown in Fig. 4e, f. In principle, the AFMR ISHE voltage would also show a peak if the temperature range extended below  $T_p$  in our resonance experiments. The same proposed coherent-to-incoherent magnon thermalization mechanism should also be applicable for the QFMR, except that both magnons are RH. In the case of the QFMR, the incoherent magnon contribution should be compared with the SSE voltage above the spin-flop transition, as is also shown in Fig. 4e, f.

In Supplementary Information Note VI, we describe the temperature dependence of the ISHE voltages at both the AFMR and the QFMR on the basis of a proposed coherent-to-incoherent thermalization mechanism that converts energy from coherent Néel order precession into phase-random thermal magnons. We capture the thermalization process phenomenologically by a temperature dependent parameter  $\eta(T)$ , as shown in equations (S8) and (S9) in Supplementary Information. We



further assume that  $\eta(T)$  scales as  $T^\alpha$ , with the exponent  $\alpha$  obtained from fitting the experimental data. The higher the temperature is, the faster incoherent magnons are generated by coherent magnons, which leads to a reduction of coherent spin pumping. On the other hand, the scattering between thermal magnons and phonons also becomes stronger at higher temperatures, which destroys incoherent magnons and eventually transfers energy into phonons. Meanwhile, the net magnon spin-current polarization, characterized by  $\xi(T)$  in equation (S15) in Supplementary Information, also decreases with increasing  $T$  (Extended Data Fig. 8). The combination of these two mechanisms results in a substantial reduction of incoherent spin current at higher  $T$  values. Whereas both coherent and incoherent contributions decay with increasing  $T$ , the incoherent contribution varies much faster than the coherent contribution, which explains the crossover in the AFMR ISHE voltage at about 45 K in our experiments. As shown in Fig. 3c, d, our theory fits the experimental data very well.

In the case of the QFMR, coherent and incoherent magnons have the same chirality, so that they generate spin currents with the same polarization; thus, there should be no sign change at any temperature. Similarly to the AFMR case, both contributions decrease with increasing  $T$ ; thus, the total spin current simply exhibits a monotonic decay with  $T$ , which agrees with the experimental data in Fig. 3d. We also notice that the situation of the QFMR is similar to what happens in ferromagnetic spin pumping, where the distinction between coherent and incoherent contributions is quite subtle (even controversial)<sup>23,24</sup>. In this regard, we believe that AFM materials provide a unique playground for studying the interplay between coherent and incoherent magnons.

In summary, we have demonstrated the generation and electrical detection of pure spin currents pumped by resonances in uniaxial AFM  $\text{Cr}_2\text{O}_3$  using sub-terahertz radiation, and have unequivocally confirmed the ISHE nature of the voltage signals. The intriguing temperature dependences of the ISHE voltages at both the AFMR and the QFMR suggest that coherent and incoherent magnons contribute to the spin current, with the latter dominating at low temperatures. Our experimental findings set the stage for further exploring spin currents in the emerging field of AFM spintronics.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1950-4>.

- Gomonay, O., Baltz, V., Brataas, A. & Tserkovnyak, Y. Antiferromagnetic spin textures and dynamics. *Nat. Phys.* **14**, 213–216 (2018).
- Kampfrath, T. et al. Coherent terahertz control of antiferromagnetic spin waves. *Nat. Photon.* **5**, 31–34 (2011).
- Tzschaschel, C. et al. Ultrafast optical excitation of coherent magnons in antiferromagnetic NiO. *Phys. Rev. B* **95**, 174407 (2017).
- Cheng, R., Xiao, J., Niu, Q. & Brataas, A. Spin pumping and spin-transfer torques in antiferromagnets. *Phys. Rev. Lett.* **113**, 057601 (2014).
- Johansen, Ø. & Brataas, A. Spin pumping and inverse spin Hall voltages from dynamical antiferromagnets. *Phys. Rev. B* **95**, 220408 (2017).
- Ross, P. et al. Antiferromagnetic resonance detected by direct current voltages in  $\text{MnF}_2/\text{Pt}$  bilayers. *J. Appl. Phys.* **118**, 233907 (2015).
- Hoffmann, A. Spin Hall effects in metals. *IEEE Trans. Magn.* **49**, 5172–5193 (2013).
- Sinova, J. et al. Spin Hall effects. *Rev. Mod. Phys.* **87**, 1213–1260 (2015).
- Li, J. et al. Observation of magnon-mediated current drag in Pt/yttrium iron garnet/Pt(Ta) trilayers. *Nat. Commun.* **7**, 10858 (2016).
- Marti, X. et al. Room-temperature antiferromagnetic memory resistor. *Nat. Mater.* **13**, 367–374 (2014).
- Wadley, P. et al. Electrical switching of an antiferromagnet. *Science* **351**, 587–590 (2016).
- Kriegner, D. et al. Multiple-stable anisotropic magnetoresistance memory in antiferromagnetic MnTe. *Nat. Commun.* **7**, 11623 (2016).
- Kittel, C. Theory of antiferromagnetic resonance. *Phys. Rev.* **82**, 565 (1951).
- Keffer, F. & Kittel, C. Theory of antiferromagnetic resonance. *Phys. Rev.* **85**, 329–337 (1952).
- Némec, P., Fiebig, M., Kampfrath, T. & Kimel, A. V. Antiferromagnetic opto-spintronics. *Nat. Phys.* **14**, 229–241 (2018).
- Baltz, V. et al. Antiferromagnetic spintronics. *Rev. Mod. Phys.* **90**, 015005 (2018).
- He, X. et al. Robust isothermal electric control of exchange bias at room temperature. *Nat. Mater.* **9**, 579–585 (2010).
- Dayhoff, E. S. Antiferromagnetic resonance in  $\text{Cr}_2\text{O}_3$ . *Phys. Rev.* **107**, 84 (1957).
- Foner, S. High-field antiferromagnetic resonance in  $\text{Cr}_2\text{O}_3$ . *Phys. Rev.* **130**, 183–197 (1963).
- Takahashi, S. et al. Pulsed electron paramagnetic resonance spectroscopy powered by a free-electron laser. *Nature* **489**, 409 (2012).
- Edwards, D. T., Zhang, Y., Glaser, S. J., Han, S. & Sherwin, M. S. Phase cycling with a 240 GHz, free electron laser-powered electron paramagnetic resonance spectrometer. *Phys. Chem. Chem. Phys.* **15**, 5707 (2013).
- Bogdanov, A. N., Zhuravlev, A. V. & Robler, U. K. Spin-flop transition in uniaxial antiferromagnets: magnetic phases, reorientation effects, and multidomain states. *Phys. Rev. B* **75**, 094425 (2007).
- Lin, W. W. & Chien, C. L. Evidence of pure spin current. Preprint at <https://arxiv.org/abs/1804.01392> (2018).
- Chen, Y. S., Lin, J. G., Huang, S. Y. & Chien, C. L. Incoherent spin pumping from YIG single crystals. *Phys. Rev. B* **99**, 220402 (2019).
- Seki, S. et al. Thermal generation of spin current in an antiferromagnet. *Phys. Rev. Lett.* **115**, 266601 (2015).
- Geprägs, S. et al. Origin of the spin Seebeck effect in compensated ferrimagnets. *Nat. Commun.* **7**, 10452 (2016).
- Cramer, J. et al. Magnon mode selective spin transport in compensated ferrimagnets. *Nano Lett.* **17**, 3334 (2017).
- Rezende, S. M., Rodríguez-Suárez, R. L. & Azevedo, A. Theory of the spin Seebeck effect in antiferromagnets. *Phys. Rev. B* **93**, 014425 (2016).
- Wu, S. M. et al. Antiferromagnetic spin Seebeck effect. *Phys. Rev. Lett.* **116**, 097204 (2016).
- Li, J. et al. Spin Seebeck effect from antiferromagnetic magnons and critical spin fluctuations in epitaxial  $\text{FeF}_2$  films. *Phys. Rev. Lett.* **122**, 217204 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

## Methods

### Characterization of Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) crystal

Cr<sub>2</sub>O<sub>3</sub> is a typical uniaxial antiferromagnet with a hexagonal crystal structure and with its magnetic easy axis along the *c* axis (Extended Data Fig. 1a). A Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) single-crystal slab with an in-plane easy axis (*c* axis) and dimensions of 5 × 5 × 1 mm<sup>3</sup> was purchased from SurfaceNet GmbH. X-ray diffraction patterns are acquired with an Empyrean X-ray diffractometer with a Cu K $\alpha$  radiation source. The clear (30 $\bar{3}$ 0) Bragg peak at a diffraction angle of  $2\theta \approx 66^\circ$  (Extended Data Fig. 1b) confirms the (10 $\bar{1}$ 0) orientation of the crystal. The absence of other peaks over a broad range of  $2\theta$  (inset of Extended Data Fig. 1b) suggests the phase purity of the crystal. Tapping-mode atomic force microscopy (Bruker Dimension, Model 5000) characterization (Extended Data Fig. 1c) is performed on the polished surface of the Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) crystal. The root-mean-square roughness over a 1  $\mu$ m × 1  $\mu$ m area is less than 0.1 nm. A clean and smooth interface is crucial for observing the spin-pumping signal in our experiments.

### EMR measurement

Continuous-wave (CW) EMR measurements are performed on a home-built CW EMR spectrometer at the Institute for Terahertz Science and Technology of the University of California, Santa Barbara. A solid-state source, which multiplies the frequency of a 15-GHz synthesizer by a factor of 16 to achieve an output frequency of 0.240 THz, produces a CW power of 55 mW (Virginia Diodes). The incident microwave power is controlled by voltage-controlled attenuation of the source and by a pair of wire-grid polarizers. The sample is secured on a sapphire piece and mounted on a Teflon stage located at the exit of an overmoded waveguide (Thomas Keating). The reflected EMR signal is measured in induction mode, where superheterodyne detection is achieved using a Schottky subharmonic mixer (Virginia Diodes) to mix the 0.240-THz signal down to 10 GHz. A home-made intermediate-frequency stage then amplifies and mixes this 10-GHz signal down to baseband. We modulate the CW source intensity with a frequency of 13.037 Hz, and the reflected signal is measured in quadrature with a pair of lock-in amplifiers (Stanford Research System SR830).

### Measurement geometry of AFM spin pumping

In our AFM spin-pumping experiments, the sample is secured on a sapphire piece and mounted on a Teflon stage at the exit of the waveguide. The *c* axis of the Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) crystal, the propagation direction of 0.240-THz microwaves and the external magnetic fields are all along the same direction.

For electrical detection, we pattern eight Pt strips and eight Ta strips alternately on the polished surface of the Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) crystal perpendicular to the *c* axis of Cr<sub>2</sub>O<sub>3</sub>. All parallel strips have the same lateral dimensions of 50  $\mu$ m × 3.5 mm and a thickness of 5 nm. Each of the independent strips produces an EMF like a battery owing to the ISHE during the spin-pumping experiments, but the ISHE voltages from the Pt and Ta strips have opposite polarities owing to their opposite spin Hall angles. To maximize the ISHE signals, we connect these individual

‘batteries’ in series—that is, positive terminal to negative terminal—to produce a large sum signal. As depicted in Extended Data Fig. 2, we have three detection geometries to probe the ISHE signals, from Pt-only, Ta-only and Pt–Ta hybrid channels. By using multiple strips in each geometry, the ISHE voltage can be effectively multiplied. In addition, a signal preamplifier (SR560) is used to amplify the voltage signals, and the gain is set to 5,000 in all voltage measurements.

### SSE measurements

To form the Cr<sub>2</sub>O<sub>3</sub>/Pt(Ta) heterostructures for the SSE measurements, 5-nm-thick Pt(Ta) is directly deposited on top of the Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) crystal by magnetron sputtering and patterned into a Hall bar with dimensions of 200  $\mu$ m × 2,740  $\mu$ m perpendicular to the *c* axis. Then, an 80-nm-thick Al<sub>2</sub>O<sub>3</sub> insulating layer is deposited by atomic layer deposition, followed by Cr(45 nm)/Au(5 nm) films covering the Hall bar channel area as a heater. In the SSE experiment, an a.c. current is applied to the Cr/Au heater to generate a vertical temperature gradient across the interface, and the double-frequency voltage response along the Pt(Ta) Hall bar channel is recorded as the spin Seebeck signal  $V_{\text{SSE}}$  using the standard lock-in technique. The frequency of the a.c. current is set at 13 Hz. An external magnetic field is applied in plane along the easy axis of the Cr<sub>2</sub>O<sub>3</sub> (that is, [0001]) or the *c* axis during the SSE measurements. We normalize the SSE voltages by the heating power to draw meaningful comparisons among different measurement conditions.

### Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

**Acknowledgements** We acknowledge discussions with S. Zhang, W. Han, I. Barsukov, Y. Liu, T. Su and Y. Liu. Work at University of California Riverside was supported through Spins and Heat in Nanoscale Electronic Systems, an Energy Frontier Research Center funded by the US Department of Energy, Office of Science, Basic Energy Sciences under award number SC0012670 (J.L., M.L., W.Y., M.A. and J.S.). The 0.240-THz measurements were performed at the Institute for Terahertz Science and Technology's (ITST) Terahertz Facilities at the University of California, Santa Barbara, which have been upgraded under NSF award number DMR-1126894. Work by C.B.W., M.K. and M.S.S. was supported by NSF MCB 1617025.

**Author contributions** J.S. conceived the experiments and supervised the project. J.L. and M.L. fabricated the devices for both the AFMR and SSE experiments with the help of W.Y. and M.A. J.L. and C.B.W. performed the AFMR experiments with the technical assistance of M.K. and N.A., under the supervision of M.S.S. R.C. developed the theoretical model and performed the data analysis with J.L. and P.W. All authors contributed to the writing of the manuscript.

**Competing interests** The authors declare no competing interests.

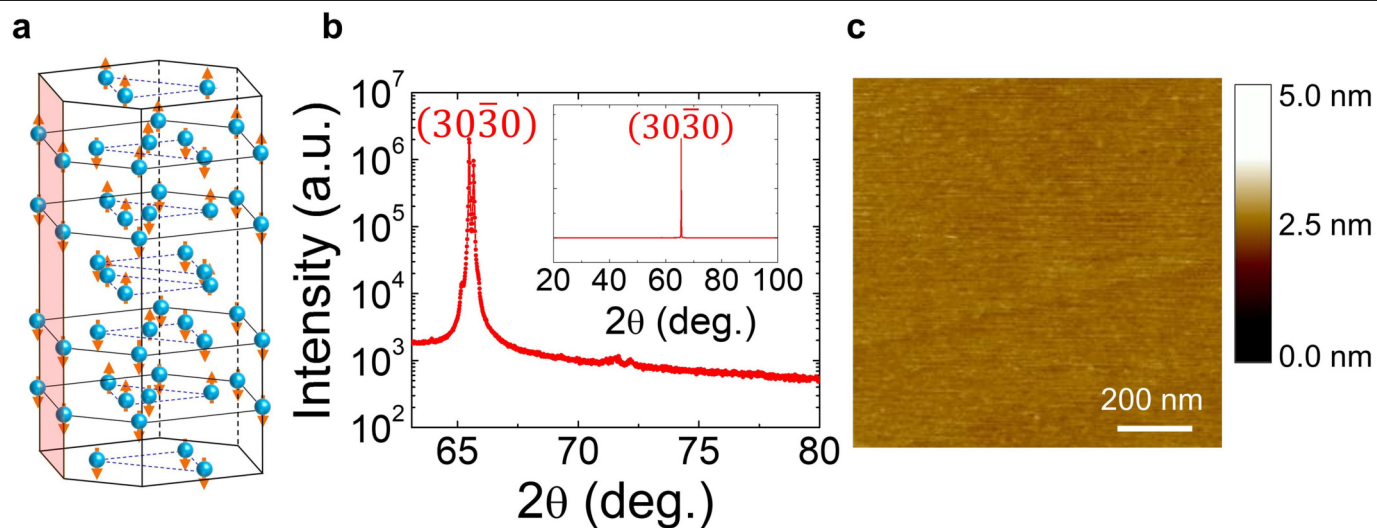
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1950-4>.

**Correspondence and requests for materials** should be addressed to J.S.

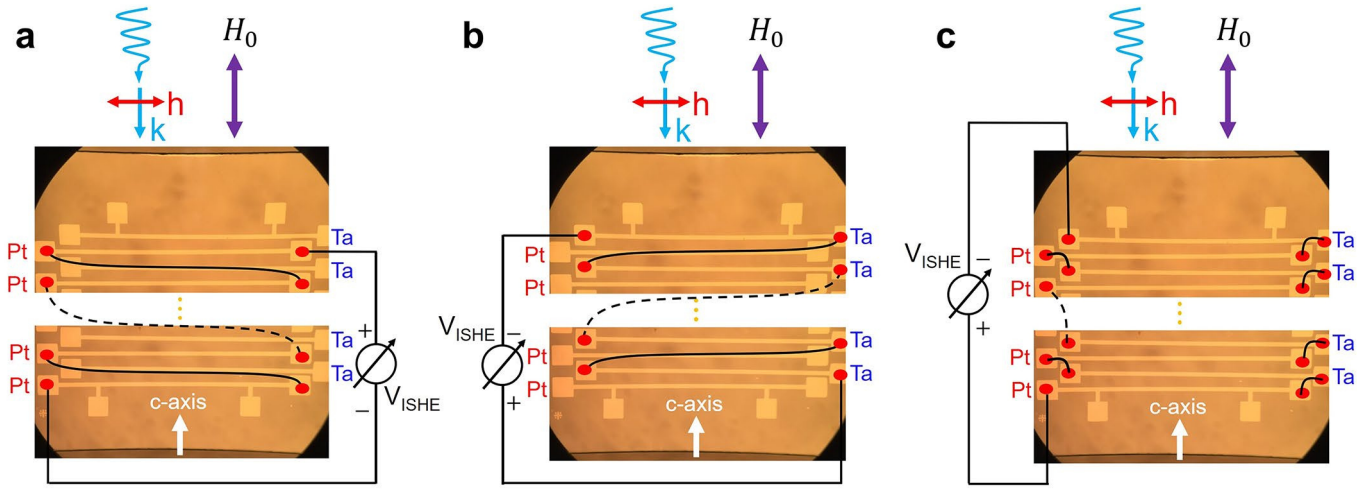
**Peer review information** *Nature* thanks Chiara Ciccarelli, Aurelien Manchon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Crystal structure and surface morphology characterization.** **a**, Crystal structure of  $\text{Cr}_2\text{O}_3$ . The symbols and arrows indicate the Cr atoms and the spins associated with them, respectively. The coloured plane is the  $(10\bar{1}0)$  plane. **b**, X-ray diffraction results of the  $\text{Cr}_2\text{O}_3$  ( $10\bar{1}0$ )

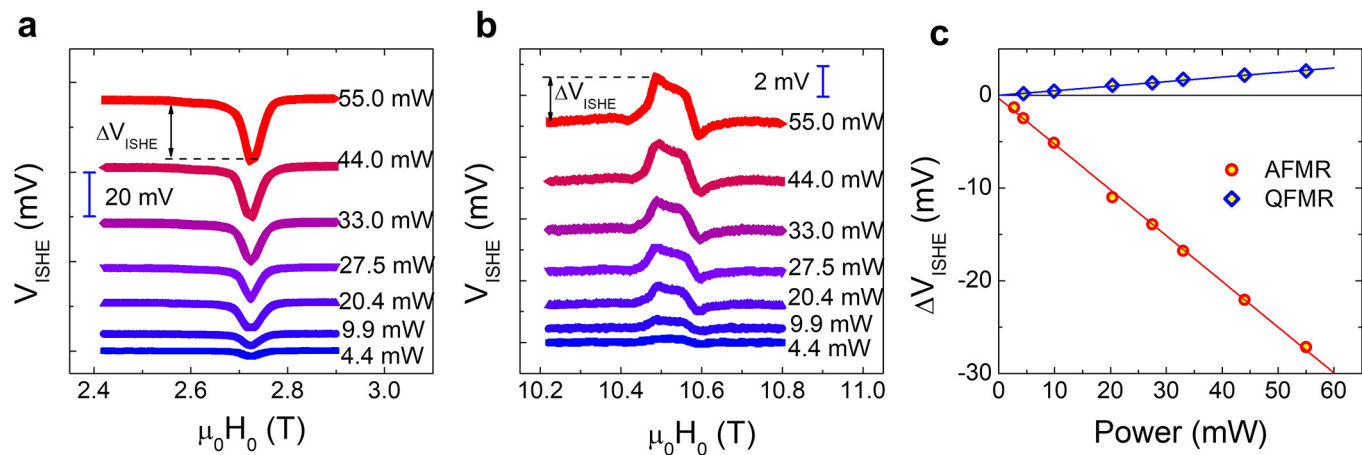
single crystal. The inset shows the X-ray diffraction results of  $\text{Cr}_2\text{O}_3$  ( $10\bar{1}0$ ) over a wide  $2\theta$  range. **c**, Atomic-force microscopy image of the polished surface of the  $\text{Cr}_2\text{O}_3$  ( $10\bar{1}0$ ) single crystal.



**Extended Data Fig. 2 | Measurement geometry of sub-terahertz spin-pumping experiments. a**, Pt channel only: only Pt strips are wire-bonded in series. **b**, Ta channel only: only Ta strips are wire-bonded in series. **c**, Pt-Ta hybrid channel. In **a–c**, black lines indicate conductive wires that connect the

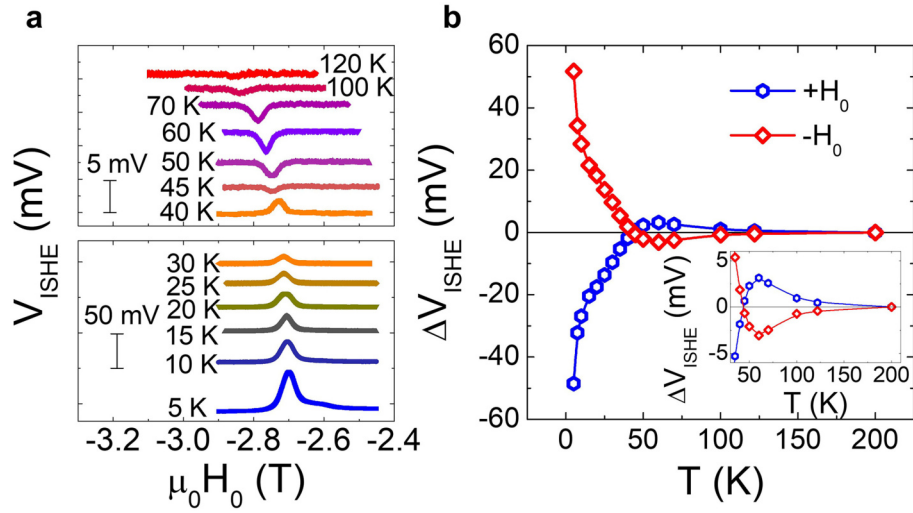
ends of the strips.  $H_0$  is an external magnetic field;  $h$  and  $k$  are the magnetic component and wavevector of the 0.240-THz microwaves, respectively.  $V_{\text{ISHE}}$  is the open-circuit voltage. The white arrows denote the  $c$  axis of  $\text{Cr}_2\text{O}_3$  ( $10\bar{1}0$ ).





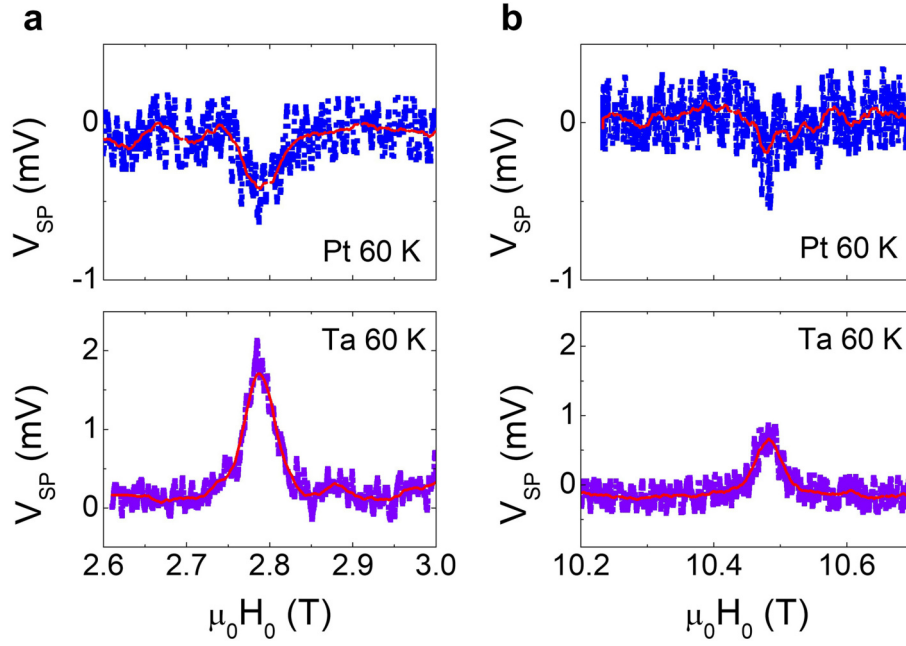
**Extended Data Fig. 3 | Linear microwave power dependence of ISHE signals at 10 K. a,** Field dependence of the ISHE signal at the AFMR for different microwave powers. **b,** Field dependence of the ISHE signal at the QFMR for

different microwave powers. **c,** Microwave power dependence of the ISHE signal magnitude at both the AFMR and the QFMR.  $\Delta V_{\text{ISHE}}$  is defined in **a** and **b**.



**Extended Data Fig. 4 | ISHE signal at the AFMR under negative external magnetic fields  $H_0$ .** **a**, ISHE signal as a function of the negative magnetic field  $H_0$  at different temperatures.  $H_0$  is along the easy axis of the Cr<sub>2</sub>O<sub>3</sub> (10 $\bar{1}$ 0) crystal.

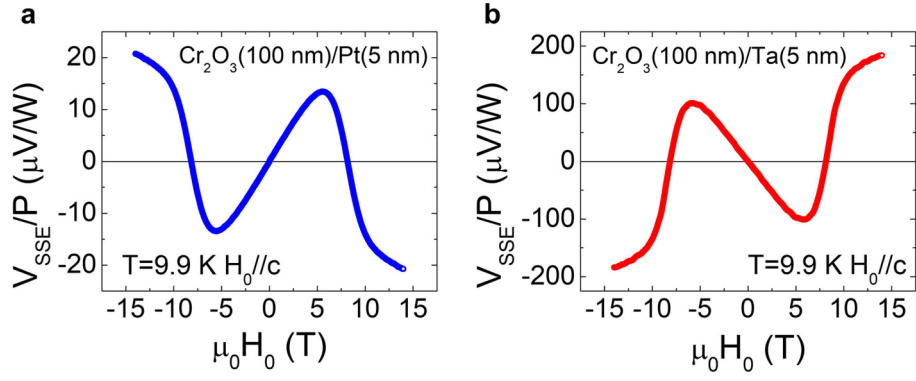
**b**, Temperature dependence of the magnitude of the ISHE signal under positive and negative magnetic fields. Inset, ISHE signal above 30 K.



**Extended Data Fig. 5 | ISHE signal from Pt- and Ta-only channels at 60 K.**

**a**, ISHE signal at the AFMR for Pt (top) and Ta (bottom) channels. **b**, ISHE signal at the QFMR for Pt (top) and Ta (bottom) channels. The red curves are smoothed ISHE signals. At the AFMR, the ISHE signals of the Pt and Ta channels at 60 K have opposite signs to that at 5 K (Fig. 2a for Pt and Fig. 2c for Ta). By

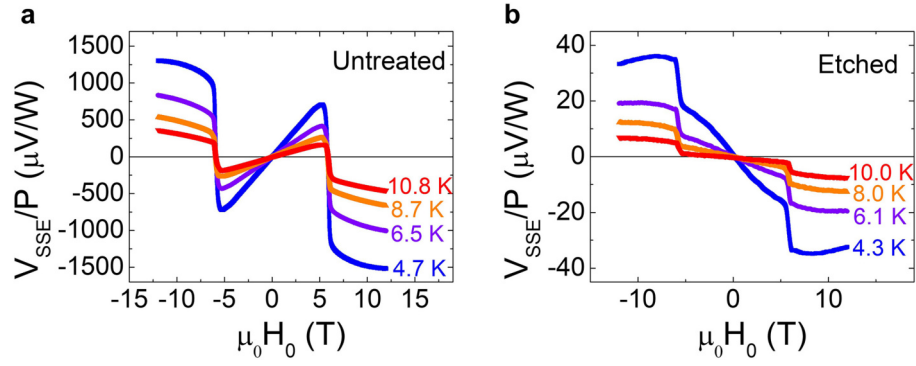
contrast, the ISHE signal for Pt and Ta at the QFMR maintains the same sign between 60 K and 5 K (Fig. 2b for Pt and Fig. 2d for Ta), which is expected because both coherent and incoherent magnons have the same chirality in the QFMR mode. At and above 60 K, the QFMR voltage signal shows a single Lorentzian peak with a slightly larger linewidth than that of the AFMR peak.



**Extended Data Fig. 6 | SSE signal at 9.9 K in  $\text{Cr}_2\text{O}_3(100 \text{ nm})/\text{Pt}$  and  $\text{Cr}_2\text{O}_3(100 \text{ nm})/\text{Ta}$  heterostructures.** **a**,  $\text{Cr}_2\text{O}_3(100 \text{ nm})/\text{Pt}$  heterostructure. **b**,  $\text{Cr}_2\text{O}_3(100 \text{ nm})/\text{Ta}$  heterostructure. The  $\text{Cr}_2\text{O}_3$  is a  $(11\bar{2}0)$ -oriented epitaxial thin film deposited on an  $\text{Al}_2\text{O}_3(11\bar{2}0)$  substrate. The magnetic field is applied

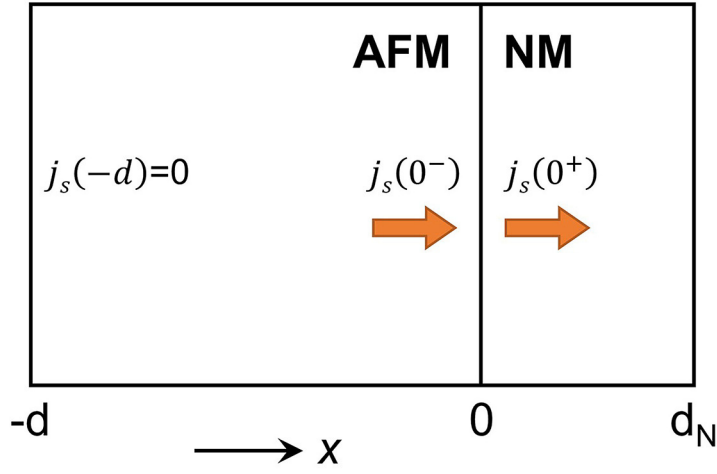
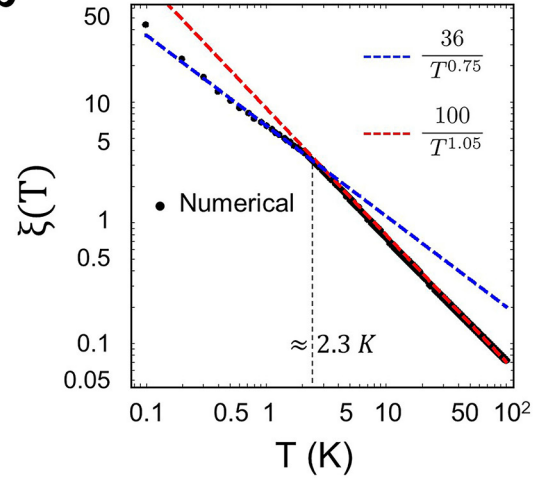
along the  $c$  axis of  $\text{Cr}_2\text{O}_3$ . The SSE signal changes sign across the spin-flop transition, which further confirms that LH magnons (dominating the SSE below the spin-flop transition) and RH magnons (dominating the SSE above the spin-flop transition) carry opposite angular momenta.





**Extended Data Fig. 7 | SSE signal in bulk  $\text{Cr}_2\text{O}_3(10\bar{1}0)/\text{Pt}$ .** **a, b**, Results are shown for bulk  $\text{Cr}_2\text{O}_3(10\bar{1}0)/\text{Pt}$  with untreated (**a**) and etched (**b**) interfaces. For the untreated sample, we anneal the crystal in air at 600 °C for 2 h using a tube furnace before the deposition of the Pt layer. For the etched sample, we first bombard the surface of the  $\text{Cr}_2\text{O}_3$  crystal with argon ions using inductively coupled plasma, and then anneal it in air at 600 °C for 2 h using a tube furnace before we deposit the Pt layer. The etching process does not affect the sign of the SSE signal above the spin-flop transition; however, it changes its sign below

the spin-flop transition. A possible reason is that the etching process may produce some uncompensated magnetic moments at the interface owing to the different sputtering yields of Cr and O atoms, and these uncompensated magnetic moments also contribute to the SSE signal by modifying the interfacial spin-mixing conductance or directly generating additional spin current. In addition, the etched sample generates a much lower SSE signal than the untreated sample under the same measurement conditions.

**a****b**

**Extended Data Fig. 8 | Schematic illustration of device used for theoretical modelling and numerical results of  $\xi(T)$ .** **a**, Schematic device geometry used to solve the spin diffusion equation of non-equilibrium incoherent magnons (equation (S9) in Supplementary Information). The bilayer structure is represented by an AFM layer and a non-magnetic (NM) metal layer of thickness

$d$  and  $d_N$ , respectively. **b**, Numerical plot and fittings of  $\xi(T)$ . Black dots are numerical calculations based on equation (S13) in Supplementary Information. Red and blue dashed lines are power-law fittings for  $T > 2.3$  K and  $T < 2.3$  K, respectively.

# Heterogeneous integration of single-crystalline complex-oxide membranes

<https://doi.org/10.1038/s41586-020-1939-z>

Received: 30 June 2019

Accepted: 4 December 2019

Published online: 5 February 2020

Hyun S. Kum<sup>1,15</sup>, Hyungwoo Lee<sup>2,15</sup>, Sungkyu Kim<sup>1,15</sup>, Shane Lindemann<sup>2,15</sup>, Wei Kong<sup>1</sup>, Kuan Qiao<sup>1</sup>, Peng Chen<sup>1</sup>, Julian Irwin<sup>3</sup>, June Hyuk Lee<sup>4</sup>, Saïen Xie<sup>5,6</sup>, Shruti Subramanian<sup>7</sup>, Jaewoo Shim<sup>1</sup>, Sang-Hoon Bae<sup>1</sup>, Chanyeol Choi<sup>8</sup>, Luigi Ranno<sup>1,9</sup>, Seungju Seo<sup>1</sup>, Sangho Lee<sup>1,9</sup>, Jackson Bauer<sup>9</sup>, Huashan Li<sup>10</sup>, Kyusang Lee<sup>11,12</sup>, Joshua A. Robinson<sup>7</sup>, Caroline A. Ross<sup>9</sup>, Darrell G. Schlom<sup>5,6</sup>, Mark S. Rzchowski<sup>3</sup>, Chang-Beom Eom<sup>2\*</sup> & Jeehwan Kim<sup>1,9,13,14\*</sup>

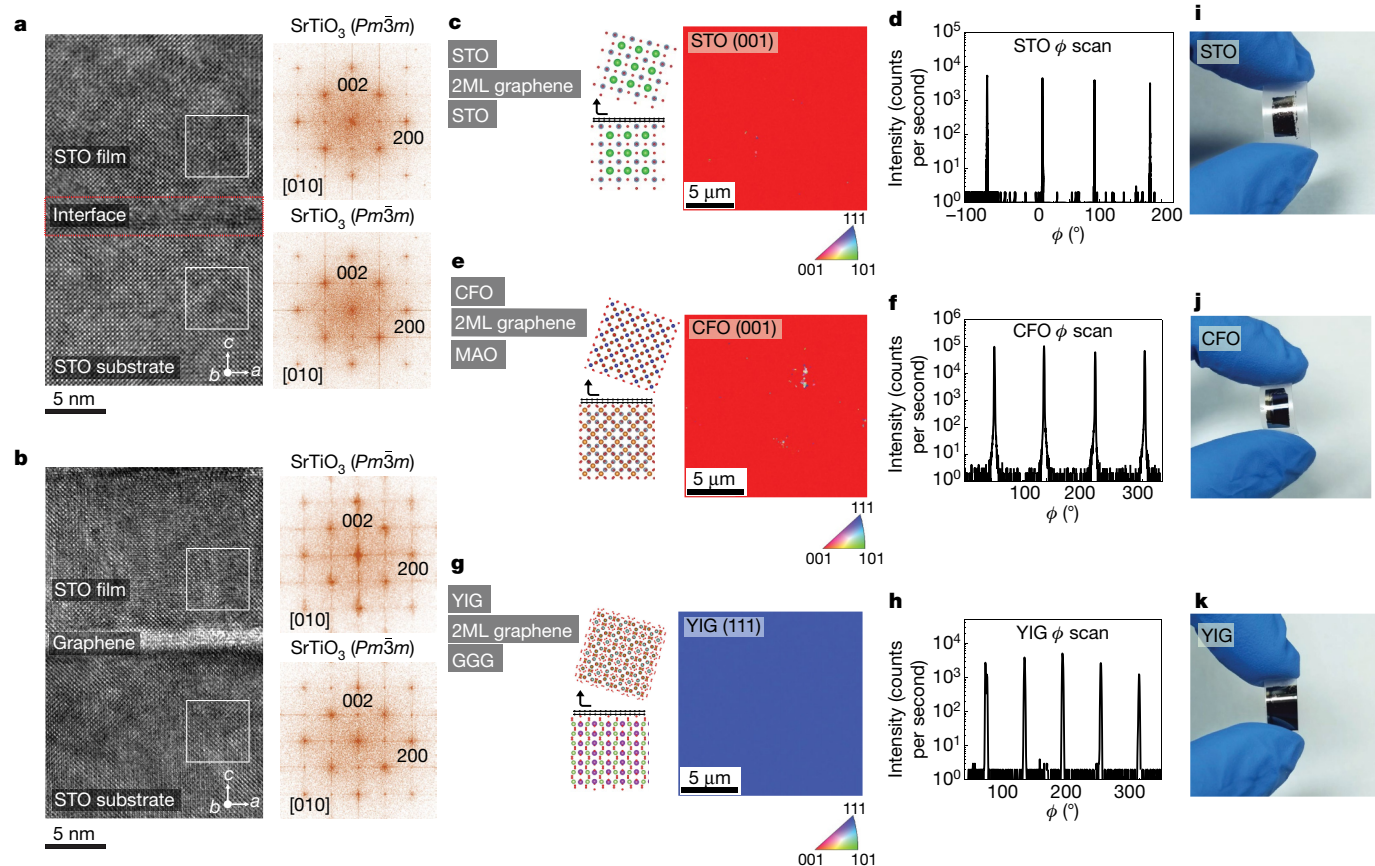
Complex-oxide materials exhibit a vast range of functional properties desirable for next-generation electronic, spintronic, magnetoelectric, neuromorphic, and energy conversion storage devices<sup>1–4</sup>. Their physical functionalities can be coupled by stacking layers of such materials to create heterostructures and can be further boosted by applying strain<sup>5–7</sup>. The predominant method for heterogeneous integration and application of strain has been through heteroepitaxy, which drastically limits the possible material combinations and the ability to integrate complex oxides with mature semiconductor technologies. Moreover, key physical properties of complex-oxide thin films, such as piezoelectricity and magnetostriction, are severely reduced by the substrate clamping effect. Here we demonstrate a universal mechanical exfoliation method of producing freestanding single-crystalline membranes made from a wide range of complex-oxide materials including perovskite, spinel and garnet crystal structures with varying crystallographic orientations. In addition, we create artificial heterostructures and hybridize their physical properties by directly stacking such freestanding membranes with different crystal structures and orientations, which is not possible using conventional methods. Our results establish a platform for stacking and coupling three-dimensional structures, akin to two-dimensional material-based heterostructures, for enhancing device functionalities<sup>8,9</sup>.

Traditionally, heterogeneous coupling and control of strain for crystalline films are carried out through heteroepitaxy on lattice-mismatched substrates<sup>10,11</sup>. Epitaxial methods, however, have fundamental limitations that prevent the unrestricted manipulation, integration and utilization of these materials. First, heteroepitaxy occurs only for different materials that have a lattice constant or crystal structure within a certain threshold. Thus, heterostructuring via epitaxy is allowed only for relatively limited material systems. Moreover, the degree of strain that can be applied to an epitaxial layer is fixed by pseudomorphic epitaxial conditions. Second, the epitaxial film is clamped by the substrate, constraining several important properties. For example, the piezoelectric and magnetostrictive responses are dampened by approximately an order of magnitude by the substrate clamping effect, reducing their sensitivity and maximum response<sup>12</sup>. Third, epitaxial growth typically requires elevated temperatures, often preventing the epitaxial integration of materials that are stable in very different

environments or are thermodynamically unstable when in contact with each other; such instability typically precludes the epitaxial integration of complex oxides with mainstream semiconductor materials. Thus, it has been extremely challenging to form heterostructures between materials with large lattice mismatch or between materials chosen solely for the desired properties they would bring to an artificial heterostructure, and it is even more challenging to unclamp epitaxial films from the substrate. Freestanding heterostructures without any limitations in crystal structures are often achieved in two-dimensional (2D) material systems by stacking ultrathin layers (a few atoms thick) of 2D materials<sup>13</sup>, and the concepts of layer transfer of single materials or of various individual devices composed of nanomaterials onto foreign substrates have been demonstrated in the past<sup>14–16</sup>. However, artificial heterostructuring of multiple single-crystalline membranes and robust physical coupling, experimentally demonstrated here, have been elusive. Although chemical lift-off of oxide materials has

<sup>1</sup>Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>2</sup>Department of Materials Science and Engineering, University of Wisconsin-Madison, Madison, WI, USA. <sup>3</sup>Department of Physics, University of Wisconsin-Madison, Madison, WI, USA. <sup>4</sup>Neutron Science Division, Korea Atomic Energy Research Institute, Daejeon, South Korea.

<sup>5</sup>Department of Materials Science and Engineering, Cornell University, Ithaca, NY, USA. <sup>6</sup>Kavli Institute at Cornell for Nanoscale Science, Ithaca, NY, USA. <sup>7</sup>Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA, USA. <sup>8</sup>Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>9</sup>Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>10</sup>Sino-French Institute for Nuclear Energy and Technology, Sun Yat-Sen University, Beijing, China. <sup>11</sup>Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, USA. <sup>12</sup>Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA, USA. <sup>13</sup>Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>14</sup>Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>15</sup>These authors contributed equally: Hyun S. Kum, Hyungwoo Lee, Sungkyu Kim, Shane Lindemann. \*e-mail: [ceom@wisc.edu](mailto:ceom@wisc.edu); [jeehwan@mit.edu](mailto:jeehwan@mit.edu)



**Fig. 1 | Epitaxial lift-off of complex-oxide membranes on graphene-coated substrates.** Cross-sectional TEM of STO film grown on a graphene-coated STO substrate without (a) and with (b) a graphene protection layer. White boxes indicate the fast Fourier transform areas, which confirm the crystallinity of each epitaxial region in comparison to the substrate. c, EBSD of the exfoliated STO membrane, confirming the single-crystalline out-of-plane (001) orientation. 2ML, two monolayers. d, Asymmetric  $\phi$  scan of the STO membrane measured by high-resolution X-ray diffraction. EBSD and

asymmetric  $\phi$  scans of CFO (100) (e, f) and YIG (111) (g, h) membranes, showing single-crystallinity with uniform out-of-plane orientation and no in-plane rotation. i–k, Photographs of the exfoliated oxide membranes (100 nm of STO (i), CFO (j) and YIG (k)) supported on thermal release tape. The strain applied to the film in the figure with a bending radius of 1.5 cm is not sufficient to crack the films<sup>42,43</sup>. The strain increases with thickness and for the 100-nm-thick films with 1.5 cm bending radius, the strain is around 0.1%, which is much smaller than the critical cracking strain of about 1%.

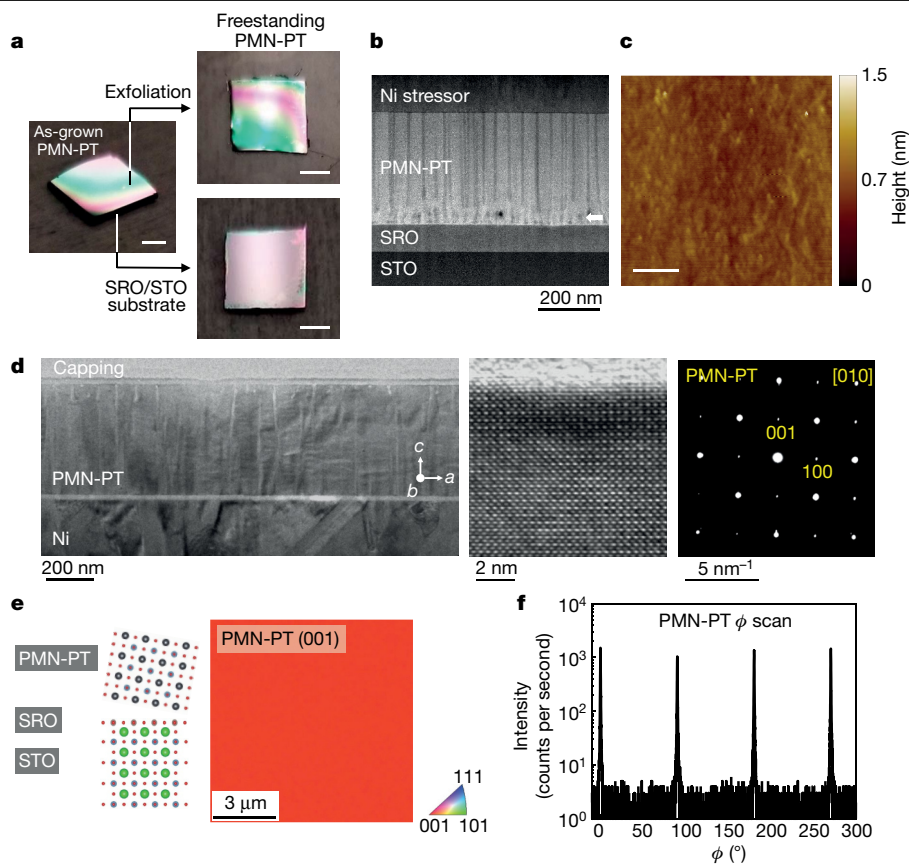
been reported<sup>17–21</sup>, this method is applicable only to a limited range of material systems owing to the lattice mismatch and etch selectivity constraints between the epitaxial layer, sacrificial layer and the substrate. Additionally, slow release rate is generally a well known shortcoming of chemical lift-off for larger substrates.

Here we prepare artificial complex-oxide heterostructure stacks using mechanical lift-off techniques, in which the epitaxial oxide films are instantly separated from weakened epitaxial interfaces to form freestanding single-crystalline membranes. These techniques can, in theory, be universally applied to prepare freestanding membranes across a broad range of crystal structures (for example, perovskite, spinel and garnet) with the potential capability of reusing the host oxide substrate. We now demonstrate freestanding membranes made from several important oxide structures including archetypal perovskite SrTiO<sub>3</sub> (STO), perovskite BaTiO<sub>3</sub> (BTO), spinel CoFe<sub>2</sub>O<sub>4</sub> (CFO), garnet Y<sub>3</sub>Fe<sub>5</sub>O<sub>12</sub> (YIG) and a perovskite of complex composition Pb(Mg<sub>1/3</sub>Nb<sub>2/3</sub>)O<sub>3</sub>–PbTiO<sub>3</sub> (PMN-PT). Single-crystalline STO, BTO, CFO and YIG were remote-epitaxially (that is, by epitaxial growth of thin films seeded by the underlying substrate through a few layers of graphene)<sup>22,23</sup> grown on graphene-coated STO, MgAl<sub>2</sub>O<sub>4</sub> (MAO) and Gd<sub>3</sub>Ga<sub>5</sub>O<sub>12</sub> (GGG) substrates, respectively, followed by mechanical exfoliation. In addition, we demonstrate single-crystalline freestanding membranes of PMN-PT that are grown via sputtering, which damages graphene<sup>24</sup>. This was enabled by discovering that SrRuO<sub>3</sub> (SRO) can provide a weak interface with PMN-PT, allowing PMN-PT films to be mechanically released precisely at the

PMN-PT/SRO interface without use of graphene. From these freestanding membranes, we fabricated various heterostructures with the goal of coupling the unique properties of the component materials by stacking them directly (that is, there is nothing between the layers). Enhanced magnetoelectric coupling has been observed by stacking magnetoelectric CFO and piezoelectric PMN-PT, because their physical properties can be greatly enhanced in freestanding form by being decoupled from the substrate. We also demonstrate magnetostatic and magnetoelastic coupling in a CFO/YIG membrane heterostructure. More importantly, we verified the electrical coupling of graphene sandwiched between freestanding CFO and YIG membranes by tracing the Fermi level shift with respect to the Dirac point of graphene. Our findings advance oxide research by allowing unrestricted integration of single-crystalline, dissimilar, complex-oxide membranes into elaborate heterostructures unattainable by epitaxy and chemical lift-off methods, which opens up opportunities to produce unprecedented three-dimensional (3D) heterostructures formed from various freestanding 2D or 3D membranes.

We first studied the growth dynamics of STO films on graphene-coated STO (001) substrates (see Supplementary Information for details of the transfer process). Our density functional theory calculation suggests that atomic potential fields can penetrate completely through bilayer graphene and partially through trilayer graphene (Extended Data Fig. 1), thus allowing successful remote epitaxy up to two monolayers of graphene interlayers. Our pulsed-laser-deposition experimental results precisely followed our prediction: single-crystalline STO films





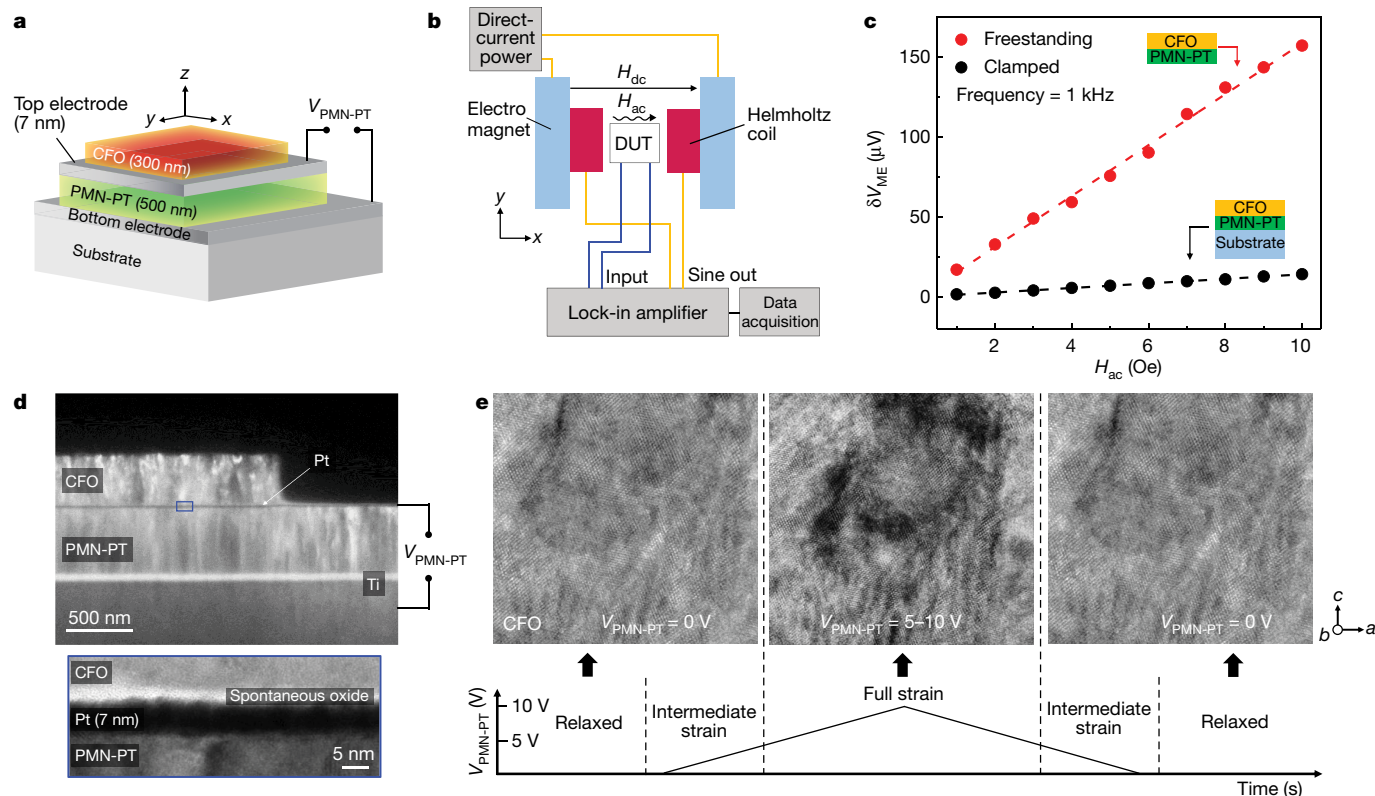
**Fig. 2 | Precise epitaxial interface separation of PMN-PT on a SRO/STO substrate.** **a**, Photograph of as-grown PMN-PT on a SRO/STO substrate (left), SRO/STO substrate after PMN-PT exfoliation (bottom right), and exfoliated PMN-PT membrane (top right). The scale bar indicates 2 mm. **b**, HAADF-STEM image of the PMN-PT/SRO/STO interface, with high-stress Ni (100 nm) deposited on top. **c**, AFM of the SRO/STO substrate surface after exfoliating the PMN-PT. The scale bar indicates 5  $\mu\text{m}$ . **d**, Cross-sectional TEM (left) of the exfoliated PMN-PT using a Ti (30 nm)/Ni (3  $\mu\text{m}$ ) metal stressor layer, verifying

the integrity of the PMN-PT crystallinity after exfoliation. Needle-like contrast areas are caused by slight variations in composition and phase that are attributable to misfit strain<sup>44</sup>. Representative high-resolution TEM (middle) and selective-area diffraction (right) images are also shown, verifying the single-crystalline nature of the exfoliated PMN-PT membrane. **e**, **f**, EBSD (**e**) and asymmetric  $\phi$  scan (**f**) of the exfoliated PMN-PT membrane, showing high-quality single-crystallinity and no in-plane rotation.

were successfully grown through bilayer graphene interlayers providing successful seeding from the STO substrates through graphene<sup>25</sup> (see Methods for epitaxy conditions). In situ high-pressure reflection high-energy electron diffraction (RHEED) during growth also showed clear intensity oscillations and crystallinity of the film during growth (Extended Data Fig. 2). In contrast to epitaxy of semiconductor compounds from groups III–V and III–nitrides on graphene where the adatoms or substrate do not react with graphene, it is crucial to avoid oxidation of graphene for epitaxy of oxides to ensure the release of remote epitaxial STO<sup>26</sup>. Initially, we were unable to exfoliate the STO films grown in a conventional oxygen overpressure owing to the graphene being etched during the nucleation stage from the substrate as evidenced by the absence of graphene in cross-sectional transmission electron microscopy (TEM) (Fig. 1a). To protect the graphene, we deposited an ultrathin STO buffer (about 5–10 nm), which is not grown in a conventional oxygen overpressure, but under vacuum ( $<5 \times 10^{-6}$  torr). By applying this buffer, we were able to preserve the graphene (Fig. 1b), resulting in successful production of freestanding STO membranes (Extended Data Fig. 3). An electron backscatter diffraction (EBSD) map of the STO film (Fig. 1c) showed (001) cubic orientation over a large area. Additionally, azimuthal X-ray diffraction  $\phi$  scanning confirmed in-plane single-crystallinity without any rotated domains (Fig. 1d). We also verified, via electron energy loss spectroscopy, that further growth of STO in an oxygen overpressure effectively corrects the oxygen stoichiometry of the entire STO film, even the region grown under vacuum (Extended Data Fig. 4). We note that because the STO substrate is also

a source of oxygen<sup>27</sup>, the exfoliation area yield of the sample with one monolayer of graphene is low compared to the exfoliation area yield of samples with two or more graphene layers (Supplementary Fig. 1). Through these findings, we concluded that two graphene layers are optimal to achieve the highest ratio of crystal quality to exfoliation yield.

As representative cases for spinel and garnet oxides, we grew spinel CFO and garnet YIG on graphene-coated MAO (001) and GGG (111) substrates, respectively. Single-crystallinity of the grown film was again verified by EBSD and high-resolution X-ray diffraction (Fig. 1e–h). The magnetization values of freestanding CFO and YIG were within a reasonable range of the bulk values, indicating the good quality of freestanding single-crystalline membranes (Supplementary Fig. 2). We carried out cross-sectional transmission electron microscope (TEM) measurements on the exfoliated CFO membrane to confirm the crystallinity at an atomic scale. Within the single-crystalline matrix, we were able to observe localized polycrystalline domains (Extended Data Fig. 5). The case is similar for other remote epitaxial membranes (see Extended Data Fig. 3). These polycrystalline domains are probably caused by regions of non-uniform graphene thickness or organic/metal residues left from graphene transfer and due to the high sticking coefficient of oxide adatoms<sup>28</sup>. Thus, the quality of the transferred graphene on the substrate determines the exfoliation area and crystallinity of the epitaxial film. Regardless, applying a thin protection layer before growing in an oxygen environment is effective for all materials explored here: all epitaxial films were successfully released from the substrate.



**Fig. 3 | Heterogeneous integration of CFO and PMN-PT membranes for strain-mediated thin-film magnetoelectrically coupled heterostructure.** **a**, Device schematic of the CFO/PMN-PT magnetoelectric device. The CFO has an area of  $4 \times 3 \text{ mm}^2$  and a thickness of 300 nm. The PMN-PT has an area of  $5 \times 5 \text{ mm}^2$  and a thickness of 500 nm. For the clamped device, the substrate is STO and the bottom electrode is SRO. For the freestanding membrane device, the substrate is PDMS and the bottom electrode is Ti. **b**, Schematic of the setup to measure the magnetic-field-induced voltage across the PMN-PT. A small alternating-current magnetic field is supplied by a Helmholtz coil and a direct-current magnetic field is provided by an electromagnet. The magnetic-field direction is applied parallel to the sample. The voltage generated across the PMN-PT membrane is measured as a function of the alternating-current magnetic-field amplitude with a lock-in amplifier. DUT stands for device under

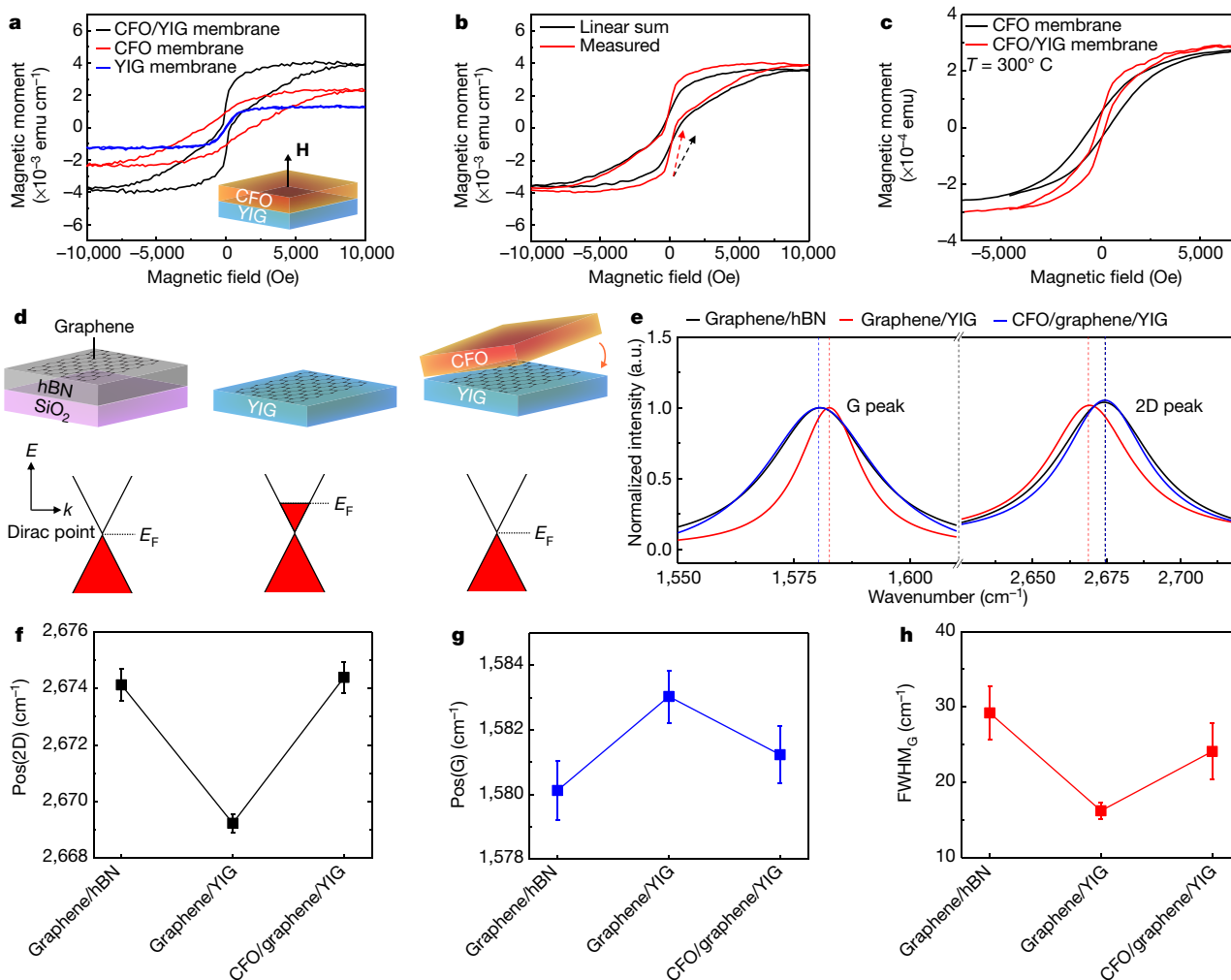
test. **c**, The voltage induced across the PMN-PT ( $\Delta V_{\text{ME}}$ ) as a function of the alternating-current magnetic field strength at a frequency of 1 kHz. The inset shows schematics of the freestanding and the clamped devices. **d**, Cross-sectional TEM image of the CFO/PMN-PT membrane heterostructure. A thin Pt (7 nm) film was deposited before transferring the CFO as the top contact to PMN-PT. The blue box shows a zoomed-in TEM of the CFO/Pt/PMN-PT interface, showing that an oxide bonding layer has formed spontaneously. **e**, In situ TEM of CFO as a function of applied voltage across the PMN-PT membrane. The strain generated by applying a voltage across the PMN-PT is transferred to the CFO layer, which induces a large change in strain contrast. The movement of the strain contrast can be more clearly seen in the Supplementary Video.

The resulting flexible STO, CFO and YIG membranes with thicknesses of 100 nm are shown in Fig. 1i–k, supported by a flexible handling tape. Moreover, we verified the compatibility of this process using molecular-beam epitaxy (MBE), another technique often used to grow single-crystalline complex-oxide films, by growing BTO on graphene-coated STO substrates and exfoliating the BTO film. Single-crystalline BTO membranes were produced, similar to the results obtained by pulsed-laser deposition (Extended Data Fig. 6).

We also discovered that the bilayer graphene interlayer not only enhances exfoliation yield, but also minimizes damage to the substrate upon peeling, thus promoting reusability of the substrates. We have successfully obtained multiple CFO freestanding membranes by reusing a single graphene-coated MAO substrate three times. The magnetic hysteresis measured from freestanding CFO using vibrating sample magnetometry was consistent throughout each exfoliation, confirming the reusability of the MAO substrate. To the best of our knowledge, our demonstration is the first to show wafer reusability for producing freestanding complex oxides, which drastically reduces production costs for applications (see Extended Data Fig. 7 and Supplementary Fig. 3 for details).

We have further broadened our mechanical exfoliation technique to oxides with more complex compositions such as PMN-PT. Single-crystalline PMN-PT films were prepared previously by sputtering<sup>24,29</sup>.

Consequently, remote epitaxy strategies cannot be applied owing to the harsh plasma overpressure that rapidly etches graphene<sup>30</sup>. We discovered that PMN-PT is weakly bonded to SRO, allowing mechanical exfoliation of PMN-PT with near-atomic precision. For this, we grew 500-nm PMN-PT/100-nm SRO epitaxial heterostructures on STO substrates by sputtering without graphene, followed by the deposition of a 3–5- $\mu\text{m}$  Ni stressor layer with a stress of around 800 MPa. Upon mechanical exfoliation, the Ni stressor provided enough strain energy to guide the crack propagation precisely at the PMN-PT/SRO interface with minimum damage to the substrate (Fig. 2a). As shown in the high-angle annular dark-field (HAADF) scanning transmission electron microscope (STEM) image in Fig. 2b, the PMN-PT/SRO interface is severely strained whereas the SRO/STO interface is pristine. After depositing high-stress Ni on PMN-PT, indications of an increased strain at the interface was observed. Our geometric phase analysis revealed that a closely spaced network of misfit dislocations (spaced about 20 nm apart) applies strain at the interface while the Ni stressor provides additional stress to the PMN-PT/SRO interface (Extended Data Fig. 8). We speculate that the resulting concentrated strain field at the PMN-PT/SRO interface provides a sufficiently weak interface to allow atomically precise crack propagation. We have reproducibly verified the precise crack propagation through the PMN-PT/SRO interface by atomic force microscopy (AFM) on the exfoliated SRO surface, which showed a root-mean-square



**Fig. 4 | Electrical, magnetostatic and magnetoelastic coupling between 3D and 2D materials.** **a**, Out-of-plane magnetic hysteresis of CFO membrane, YIG membrane, and CFO/YIG membrane heterostructures measured at room temperature. The YIG and CFO membranes individually show hysteresis loops with an in-plane easy axis resulting from shape anisotropy. The YIG has a low in-plane coercivity of  $H_{c,IP} = 10$  Oe, the CFO is magnetically harder with  $H_{c,IP} = 2,950$  Oe (Supplementary Fig. 2), and the out-of-plane loops correspond to hard-axis loops. **b**, The linear sum of the magnetic hysteresis of individual CFO and YIG membranes compared to that of the measured CFO/YIG membrane heterostructure. emu, electromagnetic unit. **c**, Magnetic hysteresis of the CFO/YIG membrane and of just the CFO membrane at 300 °C. The YIG becomes paramagnetic and only the CFO loop remains at 300 °C. When the

heterostructure membrane is cooled back down to room temperature, full recovery of the original hysteresis is observed. **d**, Schematic illustration of the electrical coupling between graphene and magnetic insulators. Graphene is undoped (the Fermi level lies at the Dirac point) on thick hBN, n-doped on YIG, and p-doped on CFO, shifting the Fermi level back to near the Dirac point when sandwiched between YIG and CFO. **e**, The Raman spectra (focused on the 2D peak and the G peak) of each structure shown in **d**. **f–h**, The shift in the 2D peak (Pos(2D)) (**f**), the shift in the G peak (Pos(G)) (**g**) and the FWHM of the G peak (FWHM<sub>G</sub>) (**h**) for graphene on hBN, YIG and sandwiched between YIG and CFO membranes, respectively. The error bars indicate the standard deviation of ten measurements from different parts of the sample.

roughness of about 2 Å (Fig. 2c). Cross-sectional TEM investigation of the exfoliated PMN-PT membranes revealed high crystalline quality as well as relieved strain at the PMN-PT surface that was bonded to the SRO before exfoliation (see Fig. 2d). Single-crystallinity over a large area has been confirmed by EBSD mapping and azimuthal  $\phi$  scans (Fig. 2e, f). In TEM, diffraction patterns taken from all imaged areas showed a single (001) orientation consistent with other characterization. We note that this technique also worked for (110) PMN-PT films grown on SRO or STO (Supplementary Fig. 4). Although details of the mechanism for the weakened interface between SRO and PMN-PT still remain to be verified, this finding further broadens the range of complex-oxide material systems that can be produced as freestanding membranes and provides opportunities to develop a graphene-free layer-release process by further exploring interface strain engineering.

Next, we fabricated heterostructures by stacking our freestanding membranes where robust mechanical coupling was observed with

high transfer yield (>90%). We first chose to stack CFO membranes onto PMN-PT membranes (see Supplementary Fig. 5 for the thin-film stacking procedure) to create a composite multiferroic (Fig. 3a). This composite allows (1) strain-mediated electric-field control of the magnetism in CFO or (2) magnetic-field-induced voltage generation across PMN-PT by virtue of the magnetostrictive and piezoelectric properties of CFO and PMN-PT, respectively.

We expect that the piezoelectric and magnetostrictive properties of these two films will be enhanced when both membranes are in their freestanding form given that they are thus free from the substrate clamping effect. PMN-PT is a material with remarkably high piezoelectric coefficient in its single-crystalline form<sup>31–33</sup>, whereas CFO has a high magnetostriction coefficient<sup>34</sup>. Thus, an enhanced strain-mediated magnetoelectric response can be expected from the stacked multiferroic heterostructure if both films are freestanding compared to when at least one of the films is clamped to the substrate. Until now,

it has only been possible to use bulk materials bonded by glue<sup>35</sup> or to grow polycrystalline films on top of piezoelectric wafers<sup>36</sup> or membranes<sup>37</sup> to realize such a hybrid structure. As expected, we observed a substantially enhanced coupling effect when both the CFO and PMN-PT were freestanding compared to that of the device where the PMN-PT was clamped to the substrate, by measuring the magnetically induced magnetoelectric coupling as shown in Fig. 3b. As shown in Fig. 3c, the freestanding CFO/PMN-PT device produces substantially larger voltage ( $\delta V_{ME}$ ) than the clamped device by more than an order of magnitude, with corresponding magnetoelectric coupling coefficients of  $477 \text{ mV cm}^{-1} \text{ Oe}^{-1}$  and  $2,675 \text{ mV cm}^{-1} \text{ Oe}^{-1}$  for the clamped and de-clamped devices, respectively. This magnetoelectric coupling coefficient is approximately an order of magnitude larger than previously reported coefficients on the same material system and comparable thickness<sup>38,39</sup>. This data indicates that strain transfer from CFO to PMN-PT is more effective for the freestanding CFO/PMN-PT heterostructure. Indeed, such excellent strain transfer between the two freestanding membranes was observed using in situ TEM (see Fig. 3d and Extended Data Fig. 9 for SEM of fabricated in situ device), showing the change in the CFO structure in response to an applied voltage ( $-10 \text{ V}$  to  $10 \text{ V}$ ) across the PMN-PT. We note that a bonding oxide layer was spontaneously formed between the CFO membrane and Pt layer for efficient strain transfer from the PMN-PT to the CFO (see bottom TEM image of Fig. 3d). The Supplementary Video shows the motion of the strain fringes in CFO in response to the strain induced by the biased PMN-PT underneath (see Fig. 3e for TEM images with or without bias showing the generation of strain fringes upon biasing). We were careful to prevent strain fringes from being generated owing to the sample flexing (see Methods for TEM sample preparation). The excellent strain transfer from freestanding PMN-PT to CFO was also verified by observing the large modulation of the magnetic hysteresis of CFO as a function of bias across PMN-PT, in contrast to the clamped device (see Extended Data Fig. 10). Thus, the CFO/PMN-PT device showcases an example of a 3D heterostructure where the functionality of each material is enhanced by stacking freestanding 3D membranes of the constituent materials.

We fabricated additional 3D complex-oxide heterostructures as well as 2D to 3D mixed heterostructures forming direct junctions to study the feasibility of new physical couplings that are not possible by conventional epitaxy. First, we observed clear magnetostatic coupling from a CFO/YIG stack. As shown in Fig. 4a, b, the measured hysteresis of the CFO/YIG heterostructure repeatedly showed a sharper reversal of its out-of-plane magnetization compared to the sum of the individual CFO and YIG loops. A signature of magnetoelastic coupling was also observed by heating the CFO/YIG heterostructure to  $300^\circ\text{C}$ , just above the Curie temperature of YIG ( $277^\circ\text{C}$ ). As shown in Fig. 4c, the loop of YIG/CFO differs from that of a single CFO layer at  $300^\circ\text{C}$  although the YIG no longer contributes a magnetic moment above its Curie temperature. We speculate that the higher thermal expansion coefficient of YIG imposes an in-plane strain on CFO, resulting in a magnetoelastic anisotropy favouring out-of-plane magnetization. Cooling to room temperature restored the two-step loop seen in Fig. 4a. These findings lay the foundations to discover physical coupling phenomena through simple stacking of these and many other functional oxides, choosing hetero-systems of interest from the huge library of freestanding membrane material sets enabled by various techniques.

This 3D heterostructuring technique not only offers great flexibility to design coupled multifunctional oxide films with enhanced performance, but also provides a platform to integrate various 3D and 2D material heterostructures with tailored functionalities to study novel interface phenomena. For example, we were able to tune the Fermi level of graphene with respect to its Dirac point by sandwiching it between YIG and CFO membranes. This was measured by tracking the 2D and G peaks of the Raman spectra of graphene, wherein contact with the YIG n-dopes the graphene and contact with CFO p-dopes the graphene, whereas graphene stays intrinsic when on thick hexagonal boron nitride

(hBN) (30 nm) (Fig. 4d)<sup>40</sup>. It is well known that the shift in the 2D and G peaks as well as the full-width at half-maximum (FWHM) change in the G peak correlates with the Fermi level<sup>41</sup>. As expected, a clear indication of n-doping was observed on graphene transferred onto a YIG membrane relative to the undoped graphene on hBN (blue-shift and red-shift of the 2D and G peaks, respectively, with a narrowing of the FWHM of the G peak). When graphene is sandwiched by transferring a CFO membrane on top of the graphene/YIG stack, graphene reverts back to a nearly undoped state similar to graphene on hBN (Fig. 4e–h). Such electrically coupled 2D and 3D mixed heterostructures provide a platform on which to study interfacial and proximity-induced physical couplings in 2D and 3D heterostructures, which so far has only been possible theoretically<sup>40</sup>. Combined with other conventional lift-off methods, it is now possible to couple and integrate an unprecedentedly broad range of functional single-crystalline membranes (III-V or III-N, complex oxides and 2D layered materials) on a single platform.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1939-z>.

- Spaldin, N. A. & Ramesh, R. Advances in magnetoelectric multiferroics. *Nat. Mater.* **18**, 203–212 (2019).
- Leung, C. M., Li, J., Viehland, D. & Zhuang, X. A review on applications of magnetoelectric composites: from heterostructural uncooled magnetic sensors, energy harvesters to highly efficient power converters. *J. Phys. D* **51**, 263002 (2018).
- Bauer, U., Przybylski, M., Kirschner, J. & Beach, G. S. D. Magnetoelectric charge trap memory. *Nano Lett.* **12**, 1437–1442 (2012).
- del Valle, J. et al. Subthreshold firing in Mott nanodevices. *Nature* **569**, 388–392 (2019).
- Schlom, D. G. et al. Strain tuning of ferroelectric thin films. *Annu. Rev. Mater. Res.* **37**, 589–626 (2007).
- Petrie, J. R., Jeon, H., Barron, S. C., Meyer, T. L. & Lee, H. N. Enhancing perovskite electrocatalysis through strain tuning of the oxygen deficiency. *J. Am. Chem. Soc.* **138**, 7252–7255 (2016).
- Ahadi, K. et al. Enhancing superconductivity in  $\text{SrTiO}_3$  films with strain. *Sci. Adv.* **5**, eaaw0120 (2019).
- Bae, S. H. et al. Integration of bulk materials with two-dimensional materials for physical coupling and applications. *Nat. Mater.* **18**, 550–560 (2019).
- Lee, H. et al. Direct observation of a two-dimensional hole gas at oxide interfaces. *Nat. Mater.* **17**, 231–236 (2018).
- Zubko, P., Gariglio, S., Gabay, M., Ghosez, P. & Triscone, J.-M. Interface physics in complex oxide heterostructures. *Annu. Rev. Condens. Matter Phys.* **2**, 141–165 (2011).
- Gan, Q., Rao, R. A., Eom, C. B., Garrett, J. L. & Lee, M. Direct measurement of strain effects on magnetic and electrical properties of epitaxial  $\text{SrRuO}_3$  thin films. *Appl. Phys. Lett.* **72**, 978–980 (1998).
- Boota, M., Houwman, E. P., Nguyen, M. D., Lanzara, G. & Rijnders, G. Effect of fabrication conditions on phase formation and properties of epitaxial  $(\text{PbMg}_{1/3}\text{Nb}_{2/3}\text{O}_3)_{0.67}/(\text{PbTiO}_3)_{0.33}$  thin films on (001)  $\text{SrTiO}_3$ . *Appl. Phys. Lett.* **6**, 055303 (2016).
- Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).
- Ahn, J. et al. Heterogeneous three-dimensional electronics by use of printed semiconductor nanomaterials. *Science* **314**, 1754–1757 (2006).
- Yim, K. H. et al. Efficient conjugated-polymer optoelectronic devices fabricated by thin-film transfer-printing technique. *Adv. Funct. Mater.* **18**, 1012–1019 (2008).
- Kum, H. et al. Epitaxial growth and layer-transfer techniques for heterogeneous integration of materials for electronic and photonic devices. *Nat. Electron.* **2**, 439–450 (2019).
- Lu, D. et al. Synthesis of freestanding single-crystal perovskite films and heterostructures by etching of sacrificial water-soluble layers. *Nat. Mater.* **15**, 1255–1260 (2016).
- Bakaul, S. R. et al. Single crystal functional oxides on silicon. *Nat. Commun.* **7**, 10547 (2016).
- Paskiewicz, D. M., Sichel-Tissot, R., Karapetrova, E., Stan, L. & Fong, D. D. Single-crystalline  $\text{SrRuO}_3$  nanomembranes: a platform for flexible oxide electronics. *Nano Lett.* **16**, 534–542 (2016).
- Ji, D. et al. Freestanding crystalline oxide perovskites down to the monolayer limit. *Nature* **570**, 87–90 (2019).
- Zhang, Y. et al. Flexible quasi-two-dimensional  $\text{CoFe}_2\text{O}_4$  epitaxial thin films for continuous strain tuning of magnetic properties. *ACS Nano* **11**, 8002–8009 (2017).
- Kim, Y. et al. Remote epitaxy through graphene enables two-dimensional material-based layer transfer. *Nature* **544**, 340–343 (2017).
- Kong, W. et al. Polarity governs atomic interaction through two-dimensional materials. *Nat. Mater.* **17**, 999–1004 (2018).
- Baek, S. H. et al. Giant piezoelectricity on Si for hyperactive MEMS. *Science* **334**, 958–961 (2011).



25. Subramanian, S. et al. Properties of synthetic epitaxial graphene/molybdenum disulfide lateral heterostructures. *Carbon* **125**, 551–556 (2017).
26. Coy-Diaz, H., Addou, R. & Batzill, M. Interface between graphene and SrTiO<sub>3</sub>(001) investigated by scanning tunneling microscopy and photoemission. *J. Phys. Chem. C* **117**, 21006–21013 (2013).
27. Schneider, C. W. et al. The origin of oxygen in oxide thin films: role of the substrate. *Appl. Phys. Lett.* **97**, 192107 (2010).
28. Willmott, P. R., Herger, R., Schlepütz, C. M., Martoccia, D. & Patterson, B. D. Energetic surface smoothing of complex metal-oxide thin films. *Phys. Rev. Lett.* **96**, 176102 (2006).
29. Brewer, A. et al. Uniform sputter deposition of high-quality epitaxial complex oxide thin films. *J. Vac. Sci. Technol. A* **35**, 060607 (2017).
30. Chen, J. et al. Self healing of defected graphene. *Appl. Phys. Lett.* **102**, 103107 (2013).
31. Li, F. et al. Ultrahigh piezoelectricity in ferroelectric ceramics by design. *Nat. Mater.* **17**, 349–354 (2018).
32. Krogstad, M. J. et al. The relation of local order to material properties in relaxor ferroelectrics. *Nat. Mater.* **17**, 718–724 (2018).
33. Li, F. et al. Giant piezoelectricity of Sm-doped Pb(Mg<sub>1/3</sub>Nb<sub>2/3</sub>)O<sub>3</sub>-PbTiO<sub>3</sub> single crystals. *Science* **1**, 264–268 (2019).
34. Bozorth, R. M., Tilden, E. F. & Williams, A. J. Anisotropy and magnetostriction of some ferrites. *Phys. Rev.* **99**, 1788 (1955).
35. Fetisov, Y. K. & Srinivasan, G. Electric field tuning characteristics of a ferrite-piezoelectric microwave resonator. *Appl. Phys. Lett.* **88**, 143503 (2006).
36. Chu, Z., PourhosseiniAsl, M. & Dong, S. Review of multi-layered magnetoelectric composite materials and devices applications. *J. Phys. D* **51**, 243001 (2018).
37. Irwin, J. et al. Magnetoelectric coupling by piezoelectric tensor design. *Sci. Rep.* **9**, 19158 (2019).
38. Feng, M., Wang, W., Zhou, Y., Li, H. & Jia, D. Influence of residual stress on magnetoelectric coupling of bilayered CoFe<sub>2</sub>O<sub>4</sub>/PMN-PT thin films. *J. Mater. Chem.* **21**, 10738–10743 (2011).
39. Wang, Z. et al. Domain rotation induced strain effect on the magnetic and magnetoelectric response in CoFe<sub>2</sub>O<sub>4</sub>/Pb(Mg,Nb)O<sub>3</sub>-PbTiO<sub>3</sub> heterostructures. *J. Appl. Phys.* **111**, 034108 (2012).
40. Hallal, A., Ibrahim, F., Yang, H., Roche, S. & Chshiev, M. Tailoring magnetic insulator proximity effects in graphene: first-principles calculations. *2D Mater.* **4**, 025074 (2017).
41. Das, A. et al. Monitoring dopants by Raman scattering in an electrochemically top-gated graphene transistor. *Nat. Nanotechnol.* **3**, 210–215 (2008).
42. Kim, J. & Xie, Y. H. Fabrication of dislocation-free tensile strained Si thin films using controllably oxidized porous Si substrates. *Appl. Phys. Lett.* **89**, 152117 (2006).
43. Suo, Z. & Hutchinson, J. W. Steady-state cracking in brittle substrates beneath adherent films. *Int. J. Solids Struct.* **25**, 1337–1353 (1989).
44. Bu, S. D. et al. Perovskite phase stabilization in epitaxial Pb(Mg<sub>1/3</sub>Nb<sub>2/3</sub>)O<sub>3</sub>-PbTiO<sub>3</sub> films by deposition onto vicinal (001) SrTiO<sub>3</sub> substrates. *Appl. Phys. Lett.* **79**, 3482–3484 (2001).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

## Methods

### Epitaxial graphene growth

Monolayer epitaxial graphene was grown via silicon sublimation from the silicon face of 6H silicon carbide (SiC (0001)) in a three-phase, hot-zone, graphite furnace (Thermal Technology LLC). In this case, a 4-inch wafer was used, and a graphite crucible was constructed to accommodate the 4-inch wafer in the furnace. The SiC was first cleaned using organic solvents (acetone, isopropyl alcohol, Nanostrip). Subsequently, the SiC is annealed in 10% hydrogen (with the rest argon) at 1,500 °C for 30 min to remove subsurface damage due to chemical and mechanical polishing. The H<sub>2</sub> was then purged from the system, and the temperature was increased to 1,800 °C for 10 min at 700 torr to form the graphene layers. This process yields low-defect-density monolayer epitaxial graphene.

### Epitaxial complex-oxide growth

**Surface preparation.** Prior to graphene transfer and growth, the STO substrate surface was dipped in buffered hydrofluoric acid for 20 s and annealed in a furnace at 1,100 °C for 6 h. AFM was measured to ensure step-and-terrace surface morphology. MAO and GGG substrates were rinsed in acetone and isopropyl alcohol for 5 min each in an ultrasonic bath with no special surface treatment.

**Pulsed laser deposition.** STO, CFO and YIG films were grown using a pulsed-laser deposition with a KrF laser energy of 400 mJ and pulse rate of 10 Hz. Commercial ceramic or bulk single-crystal targets were used. STO was grown on top of graphene-coated (100) STO substrates at a temperature of 850 °C and an oxygen flow of 20 mtorr. The initial 500 shots to the target were made without oxygen flow to protect the graphene layer on the oxide substrate for all materials. The CFO film was grown at a temperature of 400 °C and an oxygen pressure of 10 mtorr on top of a graphene-coated (100) MAO substrate. Finally, the YIG film was grown at a temperature of 700 °C and oxygen pressure of 20 mtorr on top of a graphene-coated (111) GGG substrate. After growth, the YIG film was then post-annealed at 850 °C for 2 h under an oxygen overpressure to improve crystal quality.

**Sputtering deposition.** 90° off-axis sputtering<sup>45</sup> and misaligned parallel dual planar magnetron sputtering<sup>24,29</sup> were employed to deposit epitaxial SRO and PMN-PT films, respectively. The SRO layer (100 nm) was deposited at a temperature of 600 °C and total pressure of 200 mtorr while maintaining a 3:2 ratio of Ar and O<sub>2</sub> gases. The PMN-PT layer (500 nm) was grown at a temperature of 625 °C under a total background pressure of 500 mtorr, maintaining a 17:3 ratio of Ar and O<sub>2</sub>.

**MBE deposition.** BTO films were grown by MBE in a Veeco GEN10 MBE system. Molecular beams of barium and titanium were generated using a conventional effusion cell and a Ti-Ball titanium sublimation pump, respectively. The fluxes were calibrated using RHEED intensity oscillations. Barium and titanium were co-deposited onto the substrate in an oxygen background partial pressure of  $7 \times 10^{-7}$  torr. The substrate temperature was held at 850 °C. In situ RHEED images were consistent with the growth of smooth and epitaxial thin-film surfaces during deposition.

### Graphene transfer

First, the graphene was exfoliated from its host SiC substrate by depositing Ni (~500 nm) as an adhesive/support layer. This was accomplished by first depositing a thin Ni layer using electron-beam evaporation (20 nm) to protect the graphene, followed by Ni sputtering at a chamber pressure of  $1 \times 10^{-3}$  torr and Ar flow of 9.5 standard cubic centimetres per minute (sccm). A thermal release tape (Revalpha 319Y-4M) was then used to detach the Ni layer along with the graphene. The thermal release tape/Ni/graphene stack was directly transferred onto the oxide

substrate, and the thermal release tape was released at a temperature of 120 °C. The Ni was then etched in FeCl<sub>3</sub> solution, leaving only graphene on the oxide substrate. Finally, the sample was gently rinsed in acetone and isopropyl alcohol. This process was repeated to transfer two to three layers of graphene.

### Ni stressor deposition

The Ni stressor layer was deposited using plasma sputtering, using a commercially bought Ni target with 99.99% purity. A thin Ti adhesive layer (20–80 nm) was deposited using electron-beam evaporation before depositing the Ni stressor. The Ni was sputtered at a chamber pressure of  $2 \times 10^{-3}$  torr with 9.5 sccm of Ar flow, with a growth rate of approximately 2  $\mu\text{m h}^{-1}$ .

### Characterization

**SEM, EBSD, AFM and Raman measurements.** SEM and EBSD measurements were made using a ZEISS Merlin high-resolution SEM equipped with an EBSD detector. AFM measurements were carried out using a Park NX10 AFM tool in non-contact mode. Raman spectra were obtained using a Renishaw Invia Reflex Raman confocal microscope with a laser wavelength of 532 nm, power of 1 mW and a laser spot size of 2  $\mu\text{m}$ .

**TEM measurements.** Cross-sectional TEM specimens were prepared using the focused ion beam (FEI Helios 660) technique. To prevent ion-beam damage and contamination caused by metal ions, the sample was passivated using electron-beam assisted amorphous carbon (100 nm) before using the focused ion beam. During the ion-milling process, the ion-beam energy was artificially controlled from 30 kV to 2 kV to achieve ultrathin TEM samples. Ex situ (S)TEM experiments were performed using JEOL 2010F and JEOL ARM 200CF (probe Cs-corrected) microscopes operated at 200 kV. Atomic-resolution STEM observations of epitaxial films were conducted using a JEOL ARM 200CF with a probe convergence angle of 20 mrad. A HAADF detector angle of 90–175 mrad and an annular bright-field detector angle of 11–23 mrad were used. For in situ TEM experiments, a miniature CFO/PMN-PT magnetoelectric coupled device was fabricated using the focused ion beam technique. An electron-beam assisted Pt electrode for metal probe contact was deposited onto the PMN-PT films, and the sample surface, including CFO and PMN-PT, was passivated by electron-beam induced amorphous carbon. A focused-ion-beam-cleaved specimen was connected with a metal half-grid to make the electric circuit, and this miniature device was isolated by a side-cutting method using ion milling with a low acceleration voltage of 5 kV. To remove the amorphous-carbon-assisted effect, the remaining amorphous carbon on the top of CFO was eliminated using a low-energy ion beam during the final milling stage. In situ TEM experiments were carried out using a JEOL 2010F analytical electron microscope with an acceleration voltage of 200 kV in TEM mode equipped with a biasing holder (Nanofactory Instruments AB) functionalized by a scanning tunnelling microscopy system. For electrical switching, a direct-current bias was applied inside a TEM between a sharp Pt-Ir tip operated by the scanning tunnelling microscopy function, contacting directly with the 7-nm-thick Pt layer. The TEM probe tip placement was made far from the observed CFO region (about 5  $\mu\text{m}$ ), with a relatively thick platinum contact region to minimize any effects from bending of the sample. Only negligible displacement of the sample was observed during in situ measurements, which also preclude any bending effects. Real-time high-resolution TEM videos were captured using a 2,048 pixel  $\times$  1,080 pixel resolution charge-coupled device (CCD) camera.

### Magnetoelectric device fabrication

We first transferred freestanding single-crystalline PMN-PT onto a Ti-coated polydimethylsiloxane (PDMS) substrate where Ti was used as the bottom electrode, followed by fabricating a 7-nm Pt top contact on the PMN-PT. Then, the CFO membrane was directly transferred

onto the Pt-coated PMN-PT to complete the heterostructured device. The device was annealed at 150 °C overnight to remove any moisture.

### Magnetoelectric coupling measurement

We applied a small alternating-current magnetic field at a frequency of 1 kHz on top of a direct-current magnetic field (5 kOe) in-plane across the CFO/PMN-PT device, then measured the induced voltage across the PMN-PT membrane. Voltage is generated across the PMN-PT membrane when the magnetoelastic strain in the CFO induced by the magnetic field is transferred to the PMN-PT<sup>46</sup>.

### Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

45. Eom, C. B. et al. In situ grown  $\text{YBa}_2\text{Cu}_3\text{O}_{7-d}$  thin films from single-target magnetron sputtering. *Appl. Phys. Lett.* **55**, 595–597 (1989).
46. Vopson, M. M., Fetisov, Y. K., Caruntu, G. & Srinivasan, G. Measurement techniques of the magneto-electric coupling in multiferroics. *Materials* **10**, 963 (2017).

**Acknowledgements** The team at MIT and the University of Wisconsin-Madison acknowledge support primarily by the Defense Advanced Research Projects Agency (DARPA) (award number 027049-00001, W. Carters and J. Gimlett). The work at University of Wisconsin-Madison is also supported by the Army Research Office through grant W911NF-17-1-0462. C.A.R. and J.B. acknowledge support from the SMART Center sponsored by NIST and SRC. J.A.R. and S. Subramanian acknowledge support from NSF CAREER award 1453924. J.H.L. acknowledges

support from a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (number 2018R1D1A1B07050484). The work at Cornell University is supported by the National Science Foundation (Platform for the Accelerated Realization, Analysis and Discovery of Interface Materials (PARADIM)) under Cooperative Agreement Number DMR-1539918. J.K. thanks the Masdar Institute/Khalifa University, the LG Electronics R&D Center, Amore Pacific, the LAM Research Foundation, Analogue Devices, and Rocky Mountain Vacuum Tech for general support. We are grateful to J. Li for assistance with the TEM measurements. We especially thank R. Bliem and B. Yildiz of MIT for early help in preparation of STO films.

**Author contributions** J.K. and C.B.E. conceived the idea and directed the team. H.S.K. designed and coordinated the experiments and characterization. H.S.K., H. Lee., S. Lindemann, W.K. and K.Q. performed epitaxial growth (pulsed-laser deposition and sputtering), characterization and heterogeneous integration development under the guidance of C.-B.E. and J.K. Epitaxial growth via MBE was performed by J.H.L. and S.X. under the guidance of D.G.S. Material characterization was done by H.S.K., P.C., L.R., S. Seo, C.C., S.-H.B. and K.L. Magnetoelectric coupling data analyses were performed by H.S.K, J.I. under the guidance of M.S.R. Device fabrication was carried out by H.S.K. and J.S. Magnetostatic and magnetoelastic data were analysed by H.S.K, S. Lee and J.B. under the guidance of C.A.R. The epitaxial graphene was grown by S. Subramanian under the guidance of J.A.R. Density functional theory calculations were performed by H. Li. All TEM imaging and analyses were performed by S.K. The manuscript was written by H.S.K., J.K. and C.B.E. All authors contributed to the analysis and discussion of the results leading to the manuscript.

**Competing interests** The authors declare no competing interests.

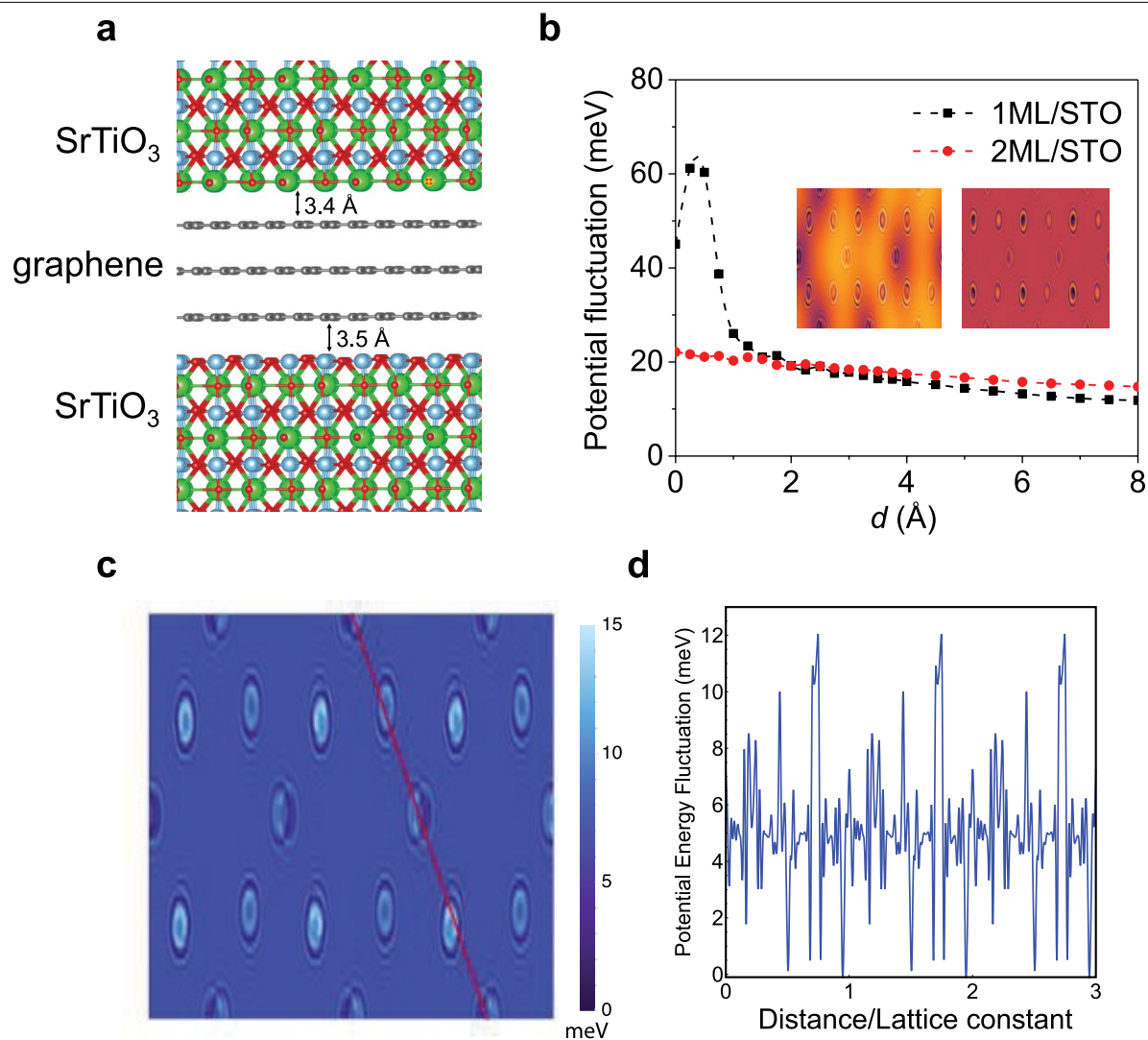
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1939-z>.

**Correspondence and requests for materials** should be addressed to C.-B.E. or J.K.

**Peer review information** *Nature* thanks Jay Switzer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

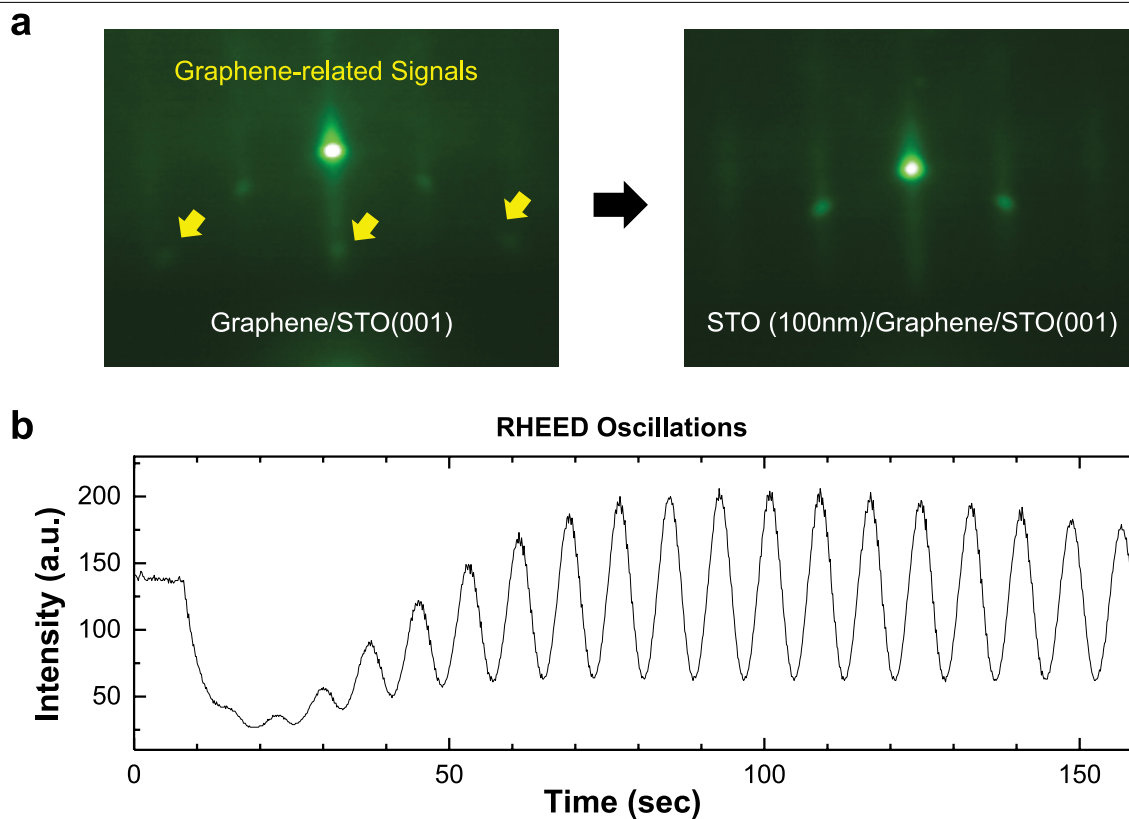
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Density functional theory simulation of substrate surface potential penetrating through graphene layers on a STO substrate.** **a**, Illustration of the simulated structure. **b**, The potential fluctuation through graphene as a function of monolayer (1ML) and bilayer (2ML) graphene thickness  $d$ . The inset shows the potential fluctuation map on the surface of

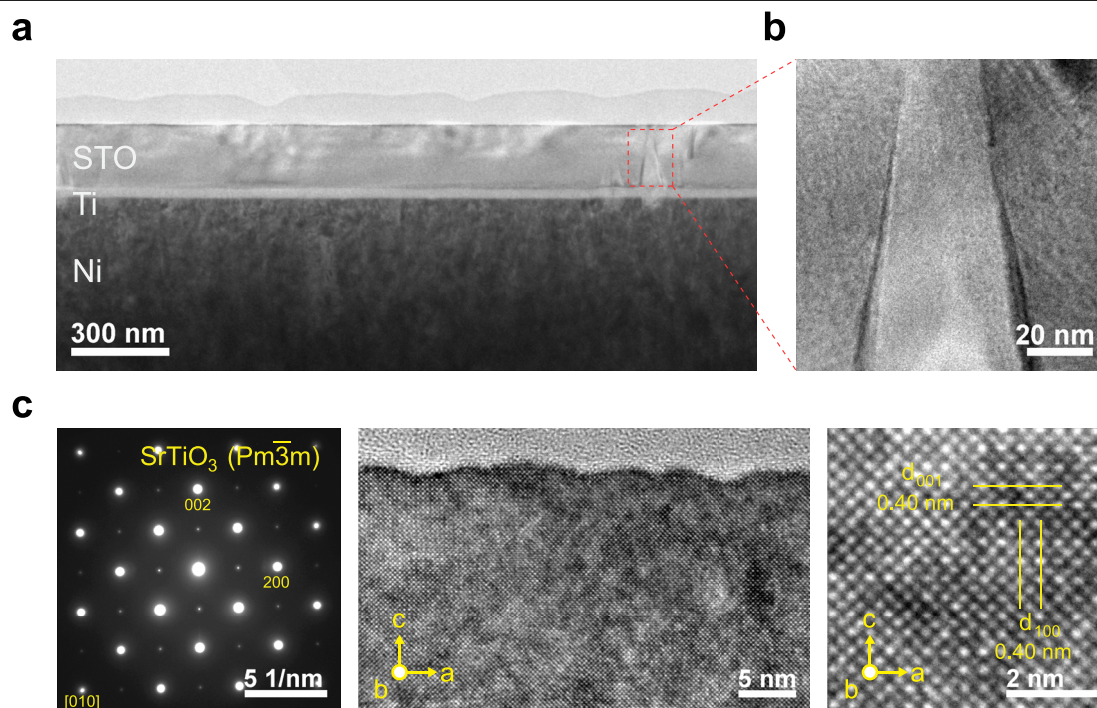
graphene-coated STO substrates for monolayer graphene (left) and bilayer graphene (right). **c**, The potential fluctuation map (colour scale) with three monolayers of graphene on top of the STO surface. **d**, Cross-sectional potential profile along the red line shown in **c**.





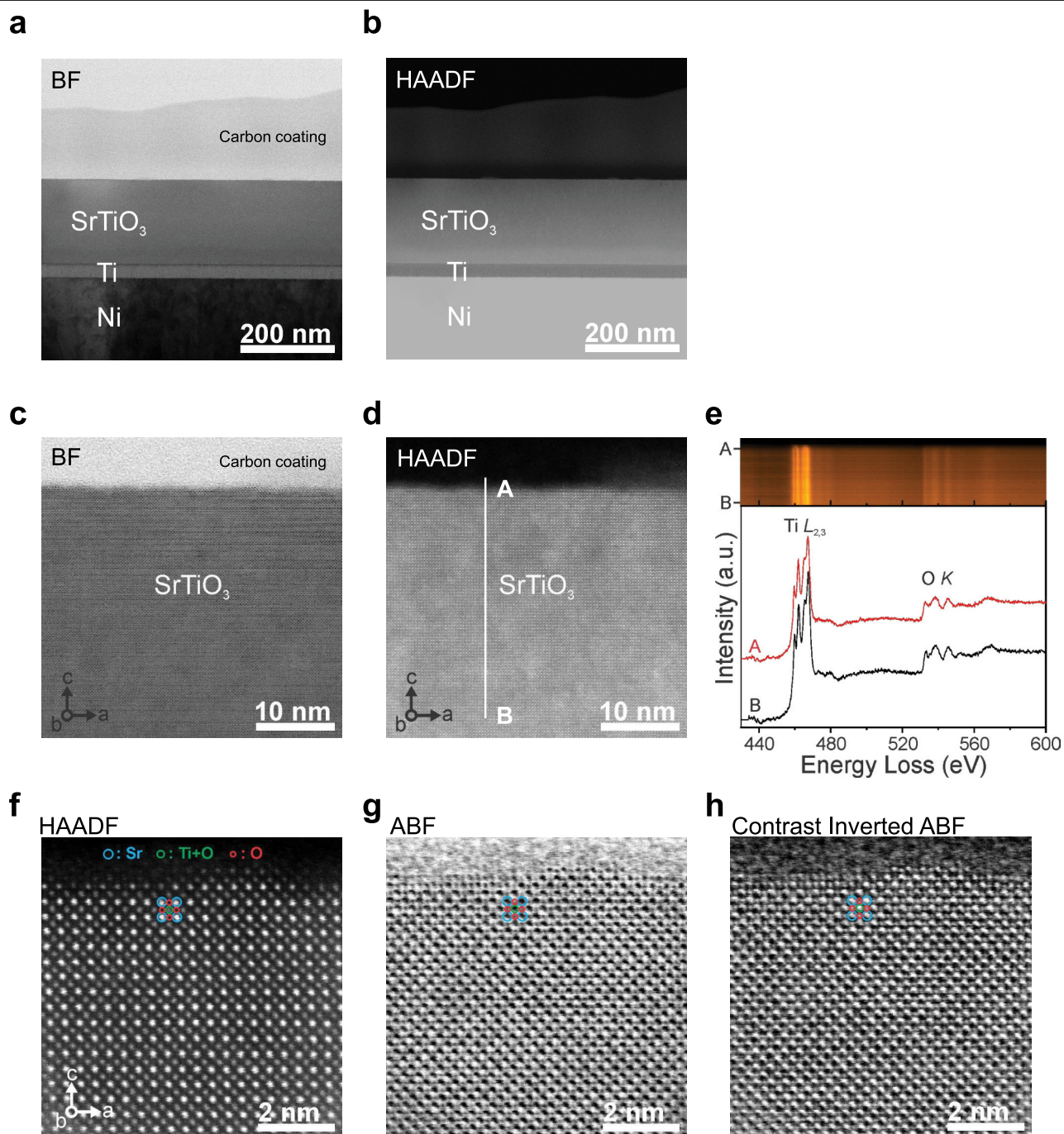
**Extended Data Fig. 2 | Pulsed-laser deposition of STO on graphene-coated STO substrate. a,** RHEED pattern during growth of STO on graphene-coated STO substrate, showing crystalline growth through the entire growth process.

The yellow arrows indicate RHEED patterns caused by the transferred graphene. **b,** RHEED oscillation during the growth of STO on the graphene-coated STO substrate.



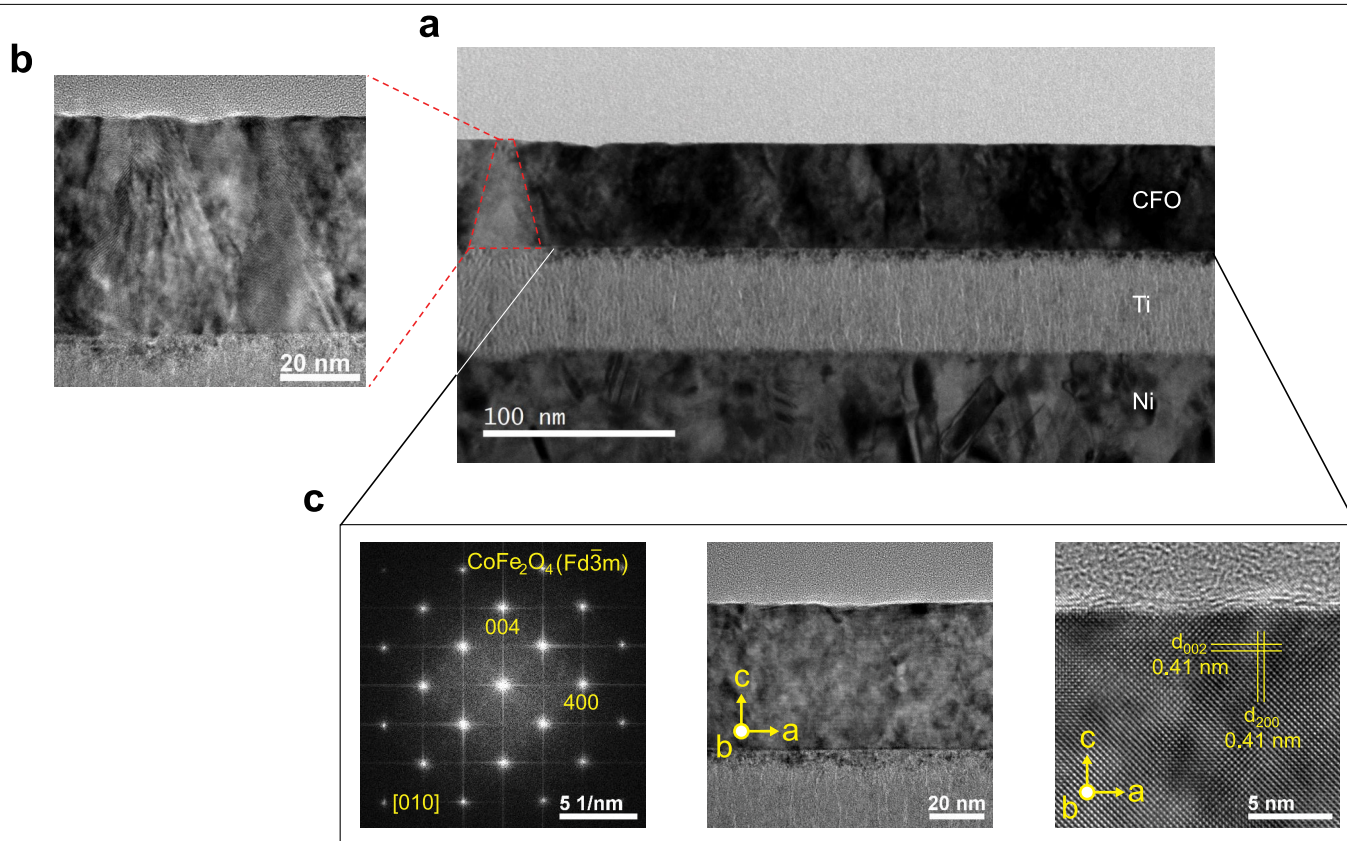
**Extended Data Fig. 3 | Cross-sectional TEM analysis of exfoliated STO membrane.** **a**, Low-magnification TEM image of STO membrane supported by the Ni stressor layer. **b**, Zoom-in of polycrystalline domains caused by residues left on graphene after transfer. **c**, High-resolution TEM image of the STO

membrane (centre), with selected area electron diffraction (left) and high-resolution TEM (right) images, confirming the overall single-crystallinity of the membrane. The single-crystalline STO membrane has a cubic structure in the  $Pm\bar{3}m$  space group with lattice distances  $d_{100}$  and  $d_{001}$  of 0.4 nm.



**Extended Data Fig. 4 | STEM analysis of the SrTiO<sub>3</sub> buffer layer grown in vacuum.** **a**, and **b**, show the bright-field (BF) and HAADF-STEM images of the exfoliated STO membrane (the same TEM sample as shown in Extended Data Fig. 3) at low magnification, respectively. **c** and **d** show higher-resolution bright-field and HAADF-STEM images of the sample, respectively. **e**, Electron energy loss spectroscopy spectra and line profile from the exfoliated surface to the bulk region (A–B in **d**), verifying that the composition of the buffer layer

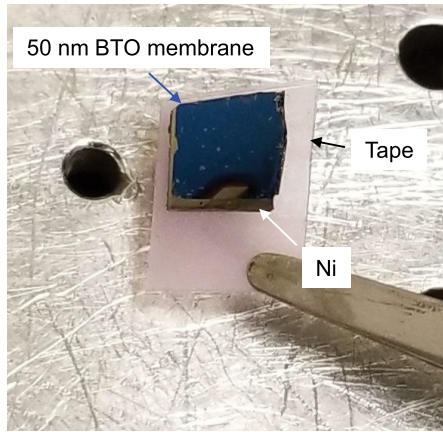
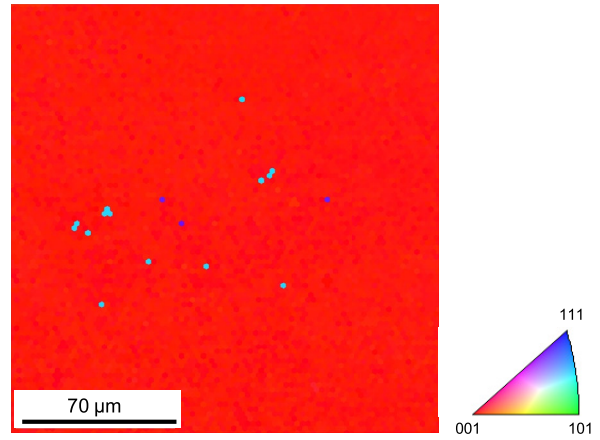
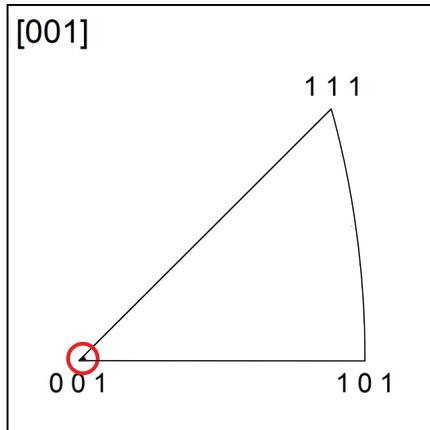
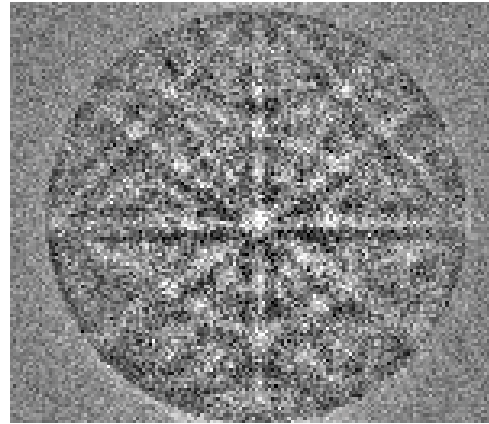
grown in vacuum is identical to the region grown under oxygen overpressure. **f**, High-resolution HAADF, showing individual atoms of the STO membrane at the exfoliation surface (the region grown in vacuum). The annular bright-field (ABF) and contrast inverted annular bright-field images (**g** and **h**, respectively) clearly show the absence of oxygen vacancies and no discernible differences are observed between the regions grown in vacuum and in oxygen.



**Extended Data Fig. 5 | Cross-sectional TEM analysis of exfoliated CFO membrane.** **a**, Cross-sectional TEM of exfoliated CFO on the Ti/Ni stressor layer. The red dotted line indicates a polycrystalline domain caused by residues left during graphene transfer. **b**, Zoomed-in TEM of the polycrystalline area marked by red dotted lines. **c**, Higher-resolution cross-sectional TEM of the

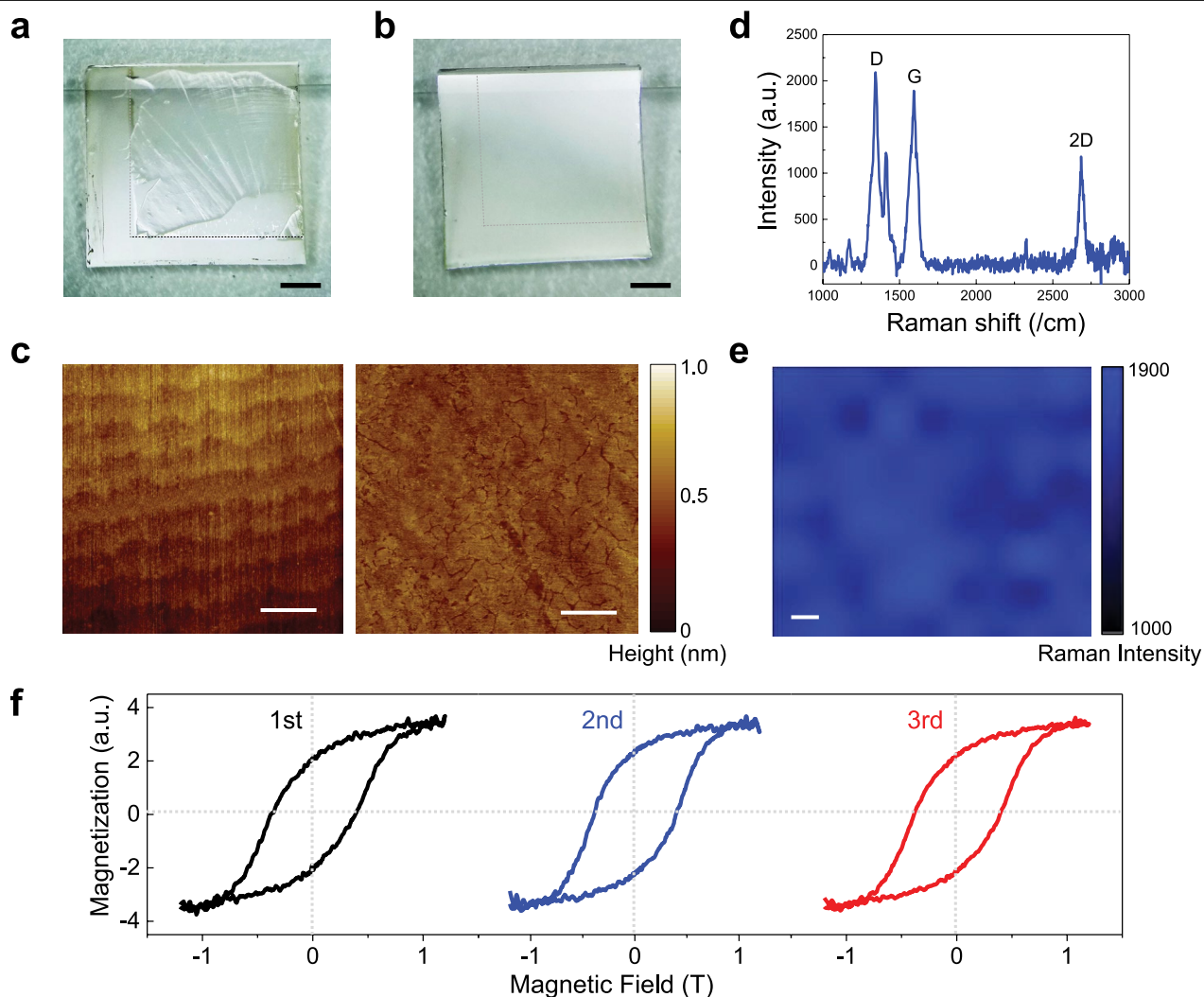
CFO film (centre), with selected area electron diffraction (left) and high-resolution TEM (right) images, confirming the overall single-crystallinity of the membrane. The single-crystalline CFO membrane has a cubic structure in the  $Fd\bar{3}m$  space group with a lattice distance  $d_{200}$  and  $d_{002}$  of 0.41 nm.



**a****b****c****d**

**Extended Data Fig. 6 | Exfoliation and characterization of  $\text{BaTiO}_3$  membrane grown via MBE.** **a**, Photograph of exfoliated BTO membrane (50 nm) grown via remote epitaxy. **b**, EBSD of the exfoliated BTO membrane showing single-

crystalline (100) orientation. **c**, The inverse-pole map of the EBSD data shown in **b**. **d**, Electron backscattering patterns (also known as Kikuchi patterns) of the BTO membrane seen over the entire area of the sample.

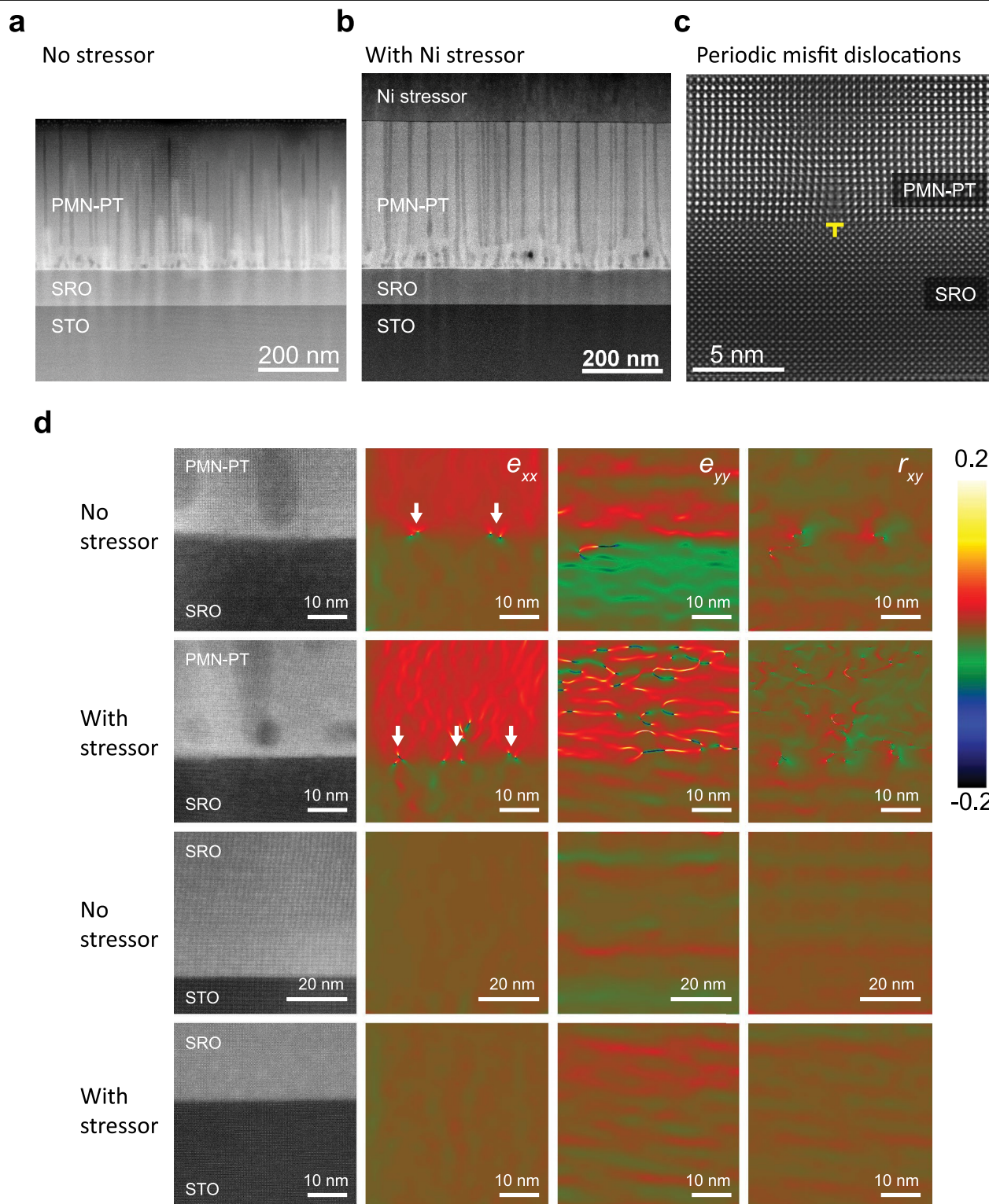


**Extended Data Fig. 7 | Reusability of a graphene-coated MAO substrate.**

**a, b**, Microscope images of a MAO substrate after exfoliating a CFO film grown on monolayer (**a**) and bilayer (**b**) graphene, where severe damage on the surface of the MAO substrate after exfoliation of CFO grown on monolayer graphene was observed, caused by crack propagation into the substrate. No evidence of damage was observed on substrates coated with bilayer graphene because the second graphene transfer covers the macroscopic defective areas of the first graphene layer. The scale bar indicates 1 mm. **c**, AFM of the pristine MAO substrate surface (left) and after one cycle of CFO exfoliation (right) with a root-mean-square roughness of approximately 5.5 Å before and after exfoliation. Scale bar indicates 1  $\mu\text{m}$ . **d, e**, Raman spectra showing the characteristic D peak,

G peak and 2D peak of graphene (**d**) and Raman intensity mapping (**e**) of the 2D peak ( $2,685\text{ cm}^{-1}$ ) of graphene on the MAO substrate after one cycle of CFO exfoliation, showing evidence that graphene is preserved on the MAO substrate after exfoliation, probably because the non-specific adhesion between graphene and CFO is weaker than that between graphene and MAO. Where this is not the case, we could etch off any graphene remaining on the substrate and re-deposit graphene before epitaxy. The scale bar indicates 10  $\mu\text{m}$ .

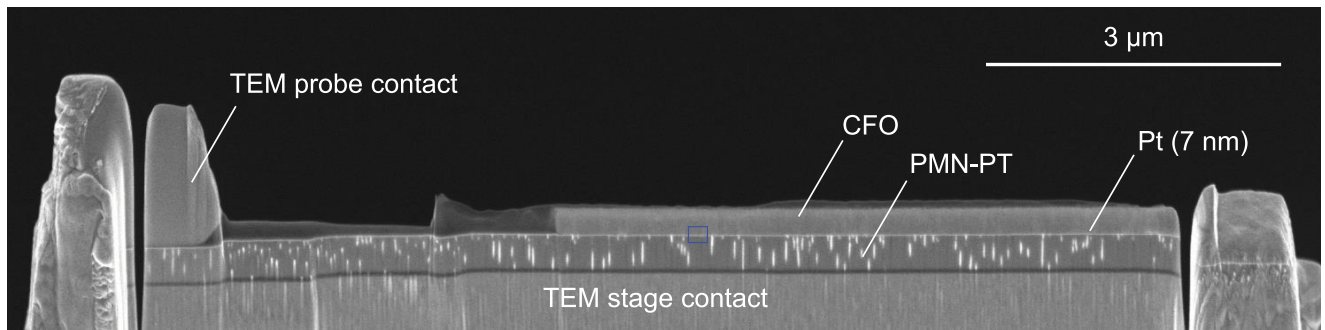
**f**, Magnetic hysteresis  $M$  of the three exfoliated CFO membranes produced on a single graphene-coated MAO substrate measured by vibrating sample magnetometry at room temperature. a.u., arbitrary units.



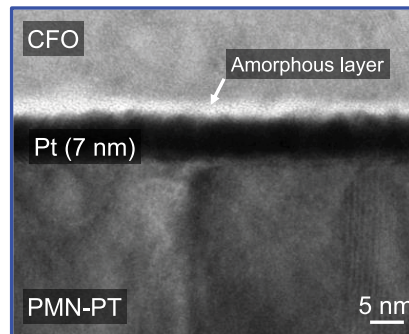
**Extended Data Fig. 8 | STEM imaging and strain analysis of the PMN-PT/SRO/STO interfaces.** **a, b**, Cross-sectional HAADF-STEM images of the PMN-PT/SRO/STO interfaces with and without a Ni stressor layer. Clear straining at the PMN-PT/SRO interface can be seen with a Ni stressor layer, whereas the SRO/STO interface remains unstrained. **c**, Atomic-resolution STEM image of one of the periodic edge dislocations observed at the PMN-PT/SRO interface.

**d**, Geometric phase analysis of the PMN-PT/SRO and SRO/STO interfaces in the  $x$  direction (2nd column),  $y$  direction (3rd column) and rotational geometry (last column) with and without the Ni stressor layer. The white arrows indicate edge dislocations. The colour scale indicates the strain fraction with reference to the SRO substrate.

a



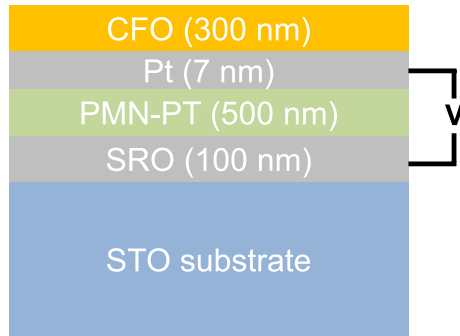
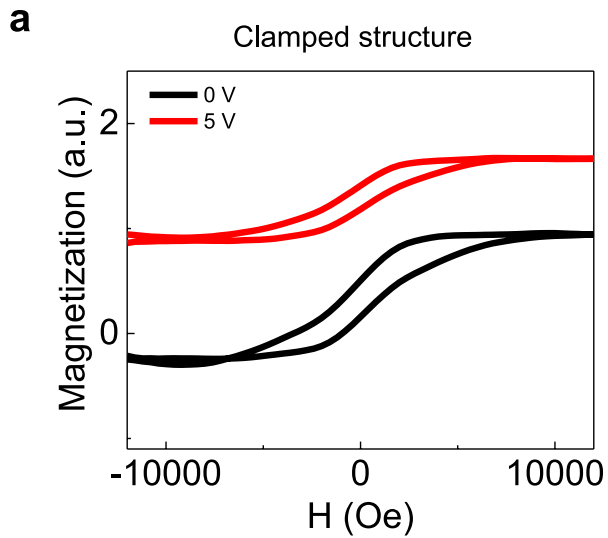
b



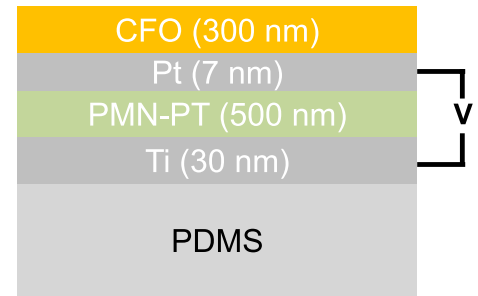
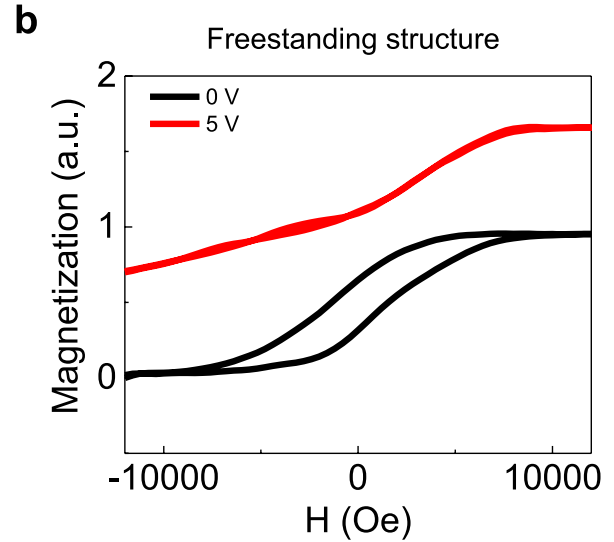
**Extended Data Fig. 9 | Description of the in situ TEM CFO/PMN-PT heterostructure device.** **a**, A cross-sectional SEM of the in situ CFO/PMN-PT magnetoelectric device. A thick Pt layer (labelled TEM probe contact) was deposited on top of the 7-nm-thick Pt layer to enable the TEM probe tip to establish electrical contact. The TEM probe contact was intentionally made

much thicker and further away from the actively observed region (distance greater than 5 μm) to prevent effects caused by bending of the sample. **b**, High-resolution TEM image of the CFO/Pt/PMN-PT interface, showing a thin amorphous oxide layer that has formed between the CFO and Pt, enabling efficient strain coupling.





**Extended Data Fig. 10 | CFO magnetic hysteresis as a function of voltage applied across PMN-PT.** CFO magnetism with a varying voltage bias across a PMN-PT measured via vibrating sample magnetometry. **a**, In the clamped structure, the PMN-PT film is grown on a SRO/STO substrate, and the CFO



membrane is transferred on top of a thin Pt layer deposited on top of PMN-PT. **b**, In the freestanding structure, the PMN-PT membrane is transferred onto a PDMS substrate after exfoliation. The rest of the stack is identical.

# Pan-cancer analysis of whole genomes

<https://doi.org/10.1038/s41586-020-1969-6>

Received: 29 July 2018

Accepted: 11 December 2019

Published online: 5 February 2020

Open access

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Cancer is driven by genetic change, and the advent of massively parallel sequencing has enabled systematic documentation of this variation at the whole-genome scale<sup>1–3</sup>. Here we report the integrative analysis of 2,658 whole-cancer genomes and their matching normal tissues across 38 tumour types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). We describe the generation of the PCAWG resource, facilitated by international data sharing using compute clouds. On average, cancer genomes contained 4–5 driver mutations when combining coding and non-coding genomic elements; however, in around 5% of cases no drivers were identified, suggesting that cancer driver discovery is not yet complete. Chromothripsis, in which many clustered structural variants arise in a single catastrophic event, is frequently an early event in tumour evolution; in acral melanoma, for example, these events precede most somatic point mutations and affect several cancer-associated genes simultaneously. Cancers with abnormal telomere maintenance often originate from tissues with low replicative activity and show several mechanisms of preventing telomere attrition to critical levels. Common and rare germline variants affect patterns of somatic mutation, including point mutations, structural variants and somatic retrotransposition. A collection of papers from the PCAWG Consortium describes non-coding mutations that drive cancer beyond those in the *TERT* promoter<sup>4</sup>; identifies new signatures of mutational processes that cause base substitutions, small insertions and deletions and structural variation<sup>5,6</sup>; analyses timings and patterns of tumour evolution<sup>7</sup>; describes the diverse transcriptional consequences of somatic mutation on splicing, expression levels, fusion genes and promoter activity<sup>8,9</sup>; and evaluates a range of more-specialized features of cancer genomes<sup>8,10–18</sup>.

Cancer is the second most-frequent cause of death worldwide, killing more than 8 million people every year; the incidence of cancer is expected to increase by more than 50% over the coming decades<sup>19,20</sup>. ‘Cancer’ is a catch-all term used to denote a set of diseases characterized by autonomous expansion and spread of a somatic clone. To achieve this behaviour, the cancer clone must co-opt multiple cellular pathways that enable it to disregard the normal constraints on cell growth, modify the local microenvironment to favour its own proliferation, invade through tissue barriers, spread to other organs and evade immune surveillance<sup>21</sup>. No single cellular program directs these behaviours. Rather, there is a large pool of potential pathogenic abnormalities from which individual cancers draw their own combinations: the commonalities of macroscopic features across tumours belie a vastly heterogeneous landscape of cellular abnormalities.

This heterogeneity arises from the stochastic nature of Darwinian evolution. There are three preconditions for Darwinian evolution: characteristics must vary within a population; this variation must be heritable from parent to offspring; and there must be competition for survival within the population. In the context of somatic cells, heritable variation arises from mutations acquired stochastically throughout life, notwithstanding additional contributions from germline and epigenetic variation. A subset of these mutations alter the cellular phenotype, and a small subset of those variants confer an advantage

on clones during the competition to escape the tight physiological controls wired into somatic cells. Mutations that provide a selective advantage to the clone are termed driver mutations, as opposed to selectively neutral passenger mutations.

Initial studies using massively parallel sequencing demonstrated the feasibility of identifying every somatic point mutation, copy-number change and structural variant (SV) in a given cancer<sup>1–3</sup>. In 2008, recognizing the opportunity that this advance in technology provided, the global cancer genomics community established the ICGC with the goal of systematically documenting the somatic mutations that drive common tumour types<sup>22</sup>.

## The pan-cancer analysis of whole genomes

The expansion of whole-genome sequencing studies from individual ICGC and TCGA working groups presented the opportunity to undertake a meta-analysis of genomic features across tumour types. To achieve this, the PCAWG Consortium was established. A Technical Working Group implemented the informatics analyses by aggregating the raw sequencing data from different working groups that studied individual tumour types, aligning the sequences to the human genome and delivering a set of high-quality somatic mutation calls for downstream analysis (Extended Data Fig. 1). Given the recent meta-analysis

A list of members and their affiliations appears in the online version of the paper and lists of working groups appear in the Supplementary Information.

## Box 1

# Online resources for data access, visualization and analysis

The PCAWG landing page (<http://docs.icgc.org/pcawg>) provides links to several data resources for interactive online browsing, analysis and download of PCAWG data and results (Supplementary Table 4).

### Direct download of PCAWG data

Aligned PCAWG read data in BAM format are also available at the European Genome Phenome Archive (EGA; <https://www.ebi.ac.uk/ega/search/site/pcawg> under accession number EGAS00001001692). In addition, all open-tier PCAWG genomics data, as well as reference datasets used for analysis, can be downloaded from the ICGC Data Portal at <http://docs.icgc.org/pcawg/data/>. Controlled-tier genomic data, including SNVs and indels that originated from TCGA projects (in VCF format) and aligned reads (in BAM format) can be downloaded using the Score (<https://www.overture.bio/>) software package, which has accelerated and secure file transfer, as well as BAM slicing facilities to selectively download defined regions of genomic alignments.

### PCAWG computational pipelines

The core alignment, somatic variant-calling, quality-control and variant consensus-generation pipelines used by PCAWG have each been packaged into portable cross-platform images using the Dockstore system<sup>84</sup> and released under an Open Source licence that enables unrestricted use and redistribution. All PCAWG Dockstore images are available to the public at <https://dockstore.org/organizations/PCAWG/collections/PCAWG>.

### ICGC Data Portal

The ICGC Data Portal<sup>85</sup> (<https://dcc.icgc.org>) serves as the main entry point for accessing PCAWG datasets with a single uniform web interface and a high-performance data-download client. This uniform interface provides users with easy access to the myriad of PCAWG sequencing data and variant calls that reside in many repositories and compute clouds worldwide. Streaming technology<sup>86</sup> provides users with high-level visualizations in real time of BAM and VCF files stored remotely on the Cancer Genome Collaboratory.

### UCSC Xena

UCSC Xena<sup>87</sup> (<https://pcawg.xenahubs.net>) visualizes all PCAWG primary results, including copy-number, gene-expression, gene-fusion and promoter-usage alterations, simple somatic mutations, large somatic structural variations, mutational signatures and phenotypic data. These open-access data are available through a public Xena hub, and consensus simple somatic mutations can be loaded to the local computer of a user via a private Xena hub. Kaplan–Meier plots, histograms, box plots, scatter plots and transcript-specific views offer additional visualization options and statistical analyses.

### The Expression Atlas

The Expression Atlas (<https://www.ebi.ac.uk/gxa/home>) contains RNA-sequencing and expression microarray data for querying gene expression across tissues, cell types, developmental stages and/or experimental conditions<sup>88</sup>. Two different views of the data are provided: summarized expression levels for each tumour type and gene expression at the level of individual samples, including reference-gene expression datasets for matching normal tissues.

### PCAWG Scout

PCAWG Scout (<http://pcawgscout.bsc.es/>) provides a framework for -omics workflow and website templating to generate on-demand, in-depth analyses of the PCAWG data that are openly available to the whole research community. Views of protected data are available that still safeguard sensitive data. Through the PCAWG Scout web interface, users can access an array of reports and visualizations that leverage on-demand bioinformatic computing infrastructure to produce results in real time, allowing users to discover trends as well as form and test hypotheses.

### Chromothripsis Explorer

Chromothripsis Explorer (<http://compbio.med.harvard.edu/chromothripsis/>) is a portal that allows structural variation in the PCAWG dataset to be explored on an individual patient basis through the use of circos plots. Patterns of chromothripsis can also be explored in aggregated formats.

of exome data from the TCGA Pan-Cancer Atlas<sup>23–25</sup>, scientific working groups concentrated their efforts on analyses best-informed by whole-genome sequencing data.

We collected genome data from 2,834 donors (Extended Data Table 1), of which 176 were excluded after quality assurance. A further 75 had minor issues that could affect some of the analyses (grey-listed donors) and 2,583 had data of optimal quality (white-listed donors) (Supplementary Table 1). Across the 2,658 white- and grey-listed donors, whole-genome sequencing data were available from 2,605 primary tumours and 173 metastases or local recurrences. Mean read coverage was 39× for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38× and 60× (Supplementary Fig. 1). RNA-sequencing data were available for 1,222 donors. The final cohort comprised 1,469 men (55%) and 1,189 women (45%), with a mean age of 56 years (range, 1–90 years) across 38 tumour types (Extended Data Table 1 and Supplementary Table 1).

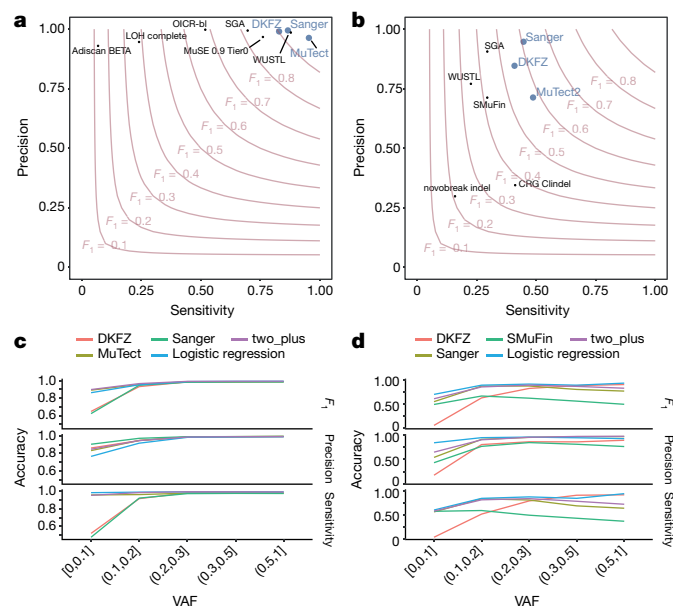
To identify somatic mutations, we analysed all 6,835 samples using a uniform set of algorithms for alignment, variant calling and quality control (Extended Data Fig. 1, Supplementary Fig. 2 and Supplementary Methods 2). We used three established pipelines to call somatic single-nucleotide variations (SNVs), small insertions and deletions (indels), copy-number alterations (CNAs) and SVs. Somatic retrotransposition events, mitochondrial DNA mutations and telomere lengths were also called by bespoke algorithms. RNA-sequencing data were uniformly

processed to call transcriptomic alterations. Germline variants identified by the three separate pipelines included single-nucleotide polymorphisms, indels, SVs and mobile-element insertions (Supplementary Table 2).

The requirement to uniformly realign and call variants on approximately 5,800 whole genomes presented considerable computational challenges, and raised ethical issues owing to the use of data from different jurisdictions (Extended Data Table 2). We used cloud computing<sup>26,27</sup> to distribute alignment and variant calling across 13 data centres on 3 continents (Supplementary Table 3). Core pipelines were packaged into Docker containers<sup>28</sup> as reproducible, stand-alone packages, which we have made available for download. Data repositories for raw and derived datasets, together with portals for data visualization and exploration, have also been created (Box 1 and Supplementary Table 4).

### Benchmarking of genetic variant calls

To benchmark mutation calling, we ran the 3 core pipelines, together with 10 additional pipelines, on 63 representative tumour–normal genome pairs (Supplementary Note 1). For 50 of these cases, we performed validation by hybridization of tumour and matched normal DNA to a custom bait set with deep sequencing<sup>29</sup>. The 3 core somatic variant-calling pipelines had individual estimates of sensitivity of 80–90% to detect a true somatic SNV called by any of the 13 pipelines; more



**Fig. 1 | Validation of variant-calling pipelines in PCAWG.** **a**, Scatter plot of estimated sensitivity and precision for somatic SNVs across individual algorithms assessed in the validation exercise across  $n = 63$  PCAWG samples. Core algorithms included in the final PCAWG call set are shown in blue. **b**, Sensitivity and precision estimates across individual algorithms for somatic indels. **c**, Accuracy (precision, sensitivity and  $F_1$  score, defined as  $2 \times \text{sensitivity} \times \text{precision} / (\text{sensitivity} + \text{precision})$ ) of somatic SNV calls across variant allele fractions (VAFs) for the core algorithms. The accuracy of two methods of combining variant calls (two-plus, which was used in the final dataset, and logistic regression) is also shown. **d**, Accuracy of indel calls across variant allele fractions.

than 95% of SNV calls made by each of the core pipelines were genuine somatic variants (Fig. 1a). For indels—a more-challenging class of variants to identify with short-read sequencing—the 3 core algorithms had individual sensitivity estimates in the range of 40–50%, with precision of 70–95% (Fig. 1b). For individual SV algorithms, we estimated precision to be in the range 80–95% for samples in the 63-sample pilot dataset.

Next, we defined a strategy to merge results from the three pipelines into one final call-set to be used for downstream scientific analyses (Methods and Supplementary Note 2). Sensitivity and precision of consensus somatic variant calls were 95% (90% confidence interval, 88–98%) and 95% (90% confidence interval, 71–99%), respectively, for SNVs (Extended Data Fig. 2). For somatic indels, sensitivity and precision were 60% (34–72%) and 91% (73–96%), respectively (Extended Data Fig. 2). Regarding somatic SVs, we estimate the sensitivity of merged calls to be 90% for true calls generated by any one pipeline; precision was estimated as 97.5%. The improvement in calling accuracy from combining different pipelines was most noticeable in variants with low variant allele fractions, which probably originate from tumour subclones (Fig. 1c, d). Germline variant calls, phased using a haplotype-reference panel, displayed a precision of more than 99% and a sensitivity of 92–98% (Supplementary Note 2).

## Analysis of PCAWG data

The uniformly generated, high-quality set of variant calls across more than 2,500 donors provided the springboard for a series of scientific working groups to explore the biology of cancer. A comprehensive suite of companion papers that describe the analyses and discoveries across these thematic areas is copublished with this paper<sup>4–18</sup> (Extended Data Table 3).

## Pan-cancer burden of somatic mutations

Across the 2,583 white-listed PCAWG donors, we called 43,778,859 somatic SNVs, 410,123 somatic multinucleotide variants, 2,418,247 somatic indels, 288,416 somatic SVs, 19,166 somatic retrotransposition events and 8,185 de novo mitochondrial DNA mutations (Supplementary Table 1). There was considerable heterogeneity in the burden of somatic mutations across patients and tumour types, with a broad correlation in mutation burden among different classes of somatic variation (Extended Data Fig. 3). Analysed at a per-patient level, this correlation held, even when considering tumours with similar purity and ploidy (Supplementary Fig. 3). Why such correlation should apply on a pan-cancer basis is unclear. It is likely that age has some role, as we observe a correlation between most classes of somatic mutation and age at diagnosis (around 190 SNVs per year,  $P = 0.02$ ; about 22 indels per year,  $P = 5 \times 10^{-5}$ ; 1.5 SVs per year,  $P < 2 \times 10^{-16}$ ; linear regression with likelihood ratio tests; Supplementary Fig. 4). Other factors are also likely to contribute to the correlations among classes of somatic mutation, as there is evidence that some DNA-repair defects can cause multiple types of somatic mutation<sup>30</sup>, and a single carcinogen can cause a range of DNA lesions<sup>31</sup>.

## Panorama of driver mutations in cancer

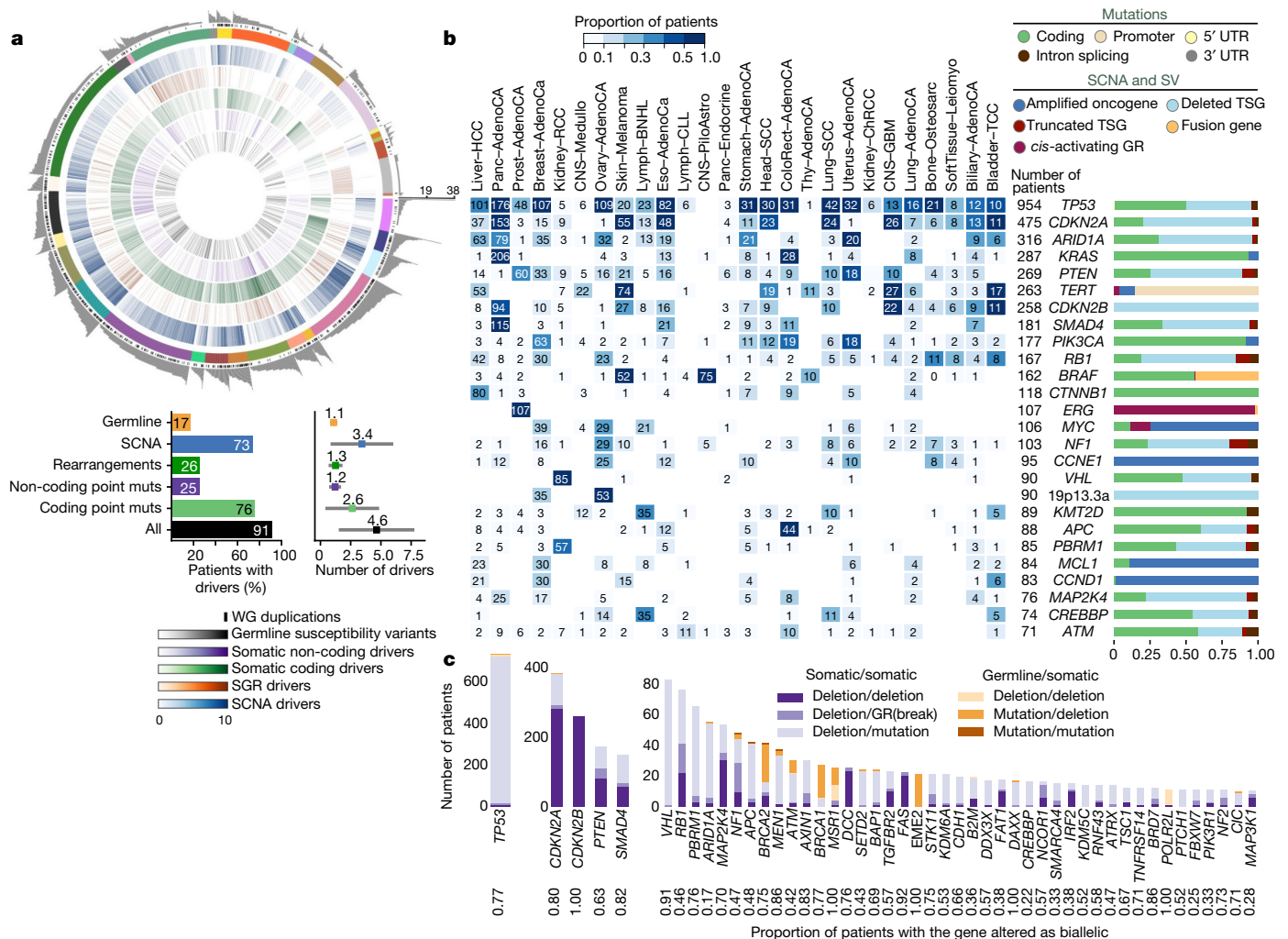
We extracted the subset of somatic mutations in PCAWG tumours that have high confidence to be driver events on the basis of current knowledge. One challenge to pinpointing the specific driver mutations in an individual tumour is that not all point mutations in recurrently mutated cancer-associated genes are drivers<sup>32</sup>. For genomic elements significantly mutated in PCAWG data, we developed a ‘rank-and-cut’ approach to identify the probable drivers (Supplementary Methods 8.1). This approach works by ranking the observed mutations in a given genomic element based on recurrence, estimated functional consequence and expected pattern of drivers in that element. We then estimate the excess burden of somatic mutations in that genomic element above that expected for the background mutation rate, and cut the ranked mutations at this level. Mutations in each element with the highest driver ranking were then assigned as probable drivers; those below the threshold will probably have arisen through chance and were assigned as probable passengers. Improvements to features that are used to rank the mutations and the methods used to measure them will contribute to further development of the rank-and-cut approach.

We also needed to account for the fact that some bona fide cancer genomic elements were not rediscovered in PCAWG data because of low statistical power. We therefore added previously known cancer-associated genes to the discovery set, creating a ‘compendium of mutational driver elements’ (Supplementary Methods 8.2). Then, using stringent rules to nominate driver point mutations that affect these genomic elements on the basis of prior knowledge<sup>33</sup>, we separated probable driver from passenger point mutations. To cover all classes of variant, we also created a compendium of known driver SVs, using analogous rules to identify which somatic CNAs and SVs are most likely to act as drivers in each tumour. For probable pathogenic germline variants, we identified all truncating germline point mutations and SVs that affect high-penetrance germline cancer-associated genes.

This analysis defined a set of mutations that we could confidently assert, based on current knowledge, drove tumorigenesis in the more than 2,500 tumours of PCAWG. We found that 91% of tumours had at least one identified driver mutation, with an average of 4.6 drivers per tumour identified, showing extensive variation across cancer types (Fig. 2a). For coding point mutations, the average was 2.6 drivers per tumour, similar to numbers estimated in known cancer-associated genes in tumours in the TCGA using analogous approaches<sup>32</sup>.

To address the frequency of non-coding driver point mutations, we combined promoters and enhancers that are known targets of





**Fig. 2 | Panorama of driver mutations in PCAWG.** **a**, Top, putative driver mutations in PCAWG, represented as a circos plot. Each sector represents a tumour in the cohort. From the periphery to the centre of the plot the concentric rings represent: (1) the total number of driver alterations; (2) the presence of whole-genome (WG) duplication; (3) the tumour type; (4) the number of driver CNAs; (5) the number of driver genomic rearrangements; (6) driver coding point mutations; (7) driver non-coding point mutations; and (8) pathogenic germline variants. Bottom, snapshots of the panorama of driver mutations. The horizontal bar plot (left) represents the proportion of patients with different types of drivers. The dot plot (right) represents the mean number of each type of driver mutation across tumours with at least one event (the square dot) and the standard deviation (grey whiskers), based on  $n = 2,583$

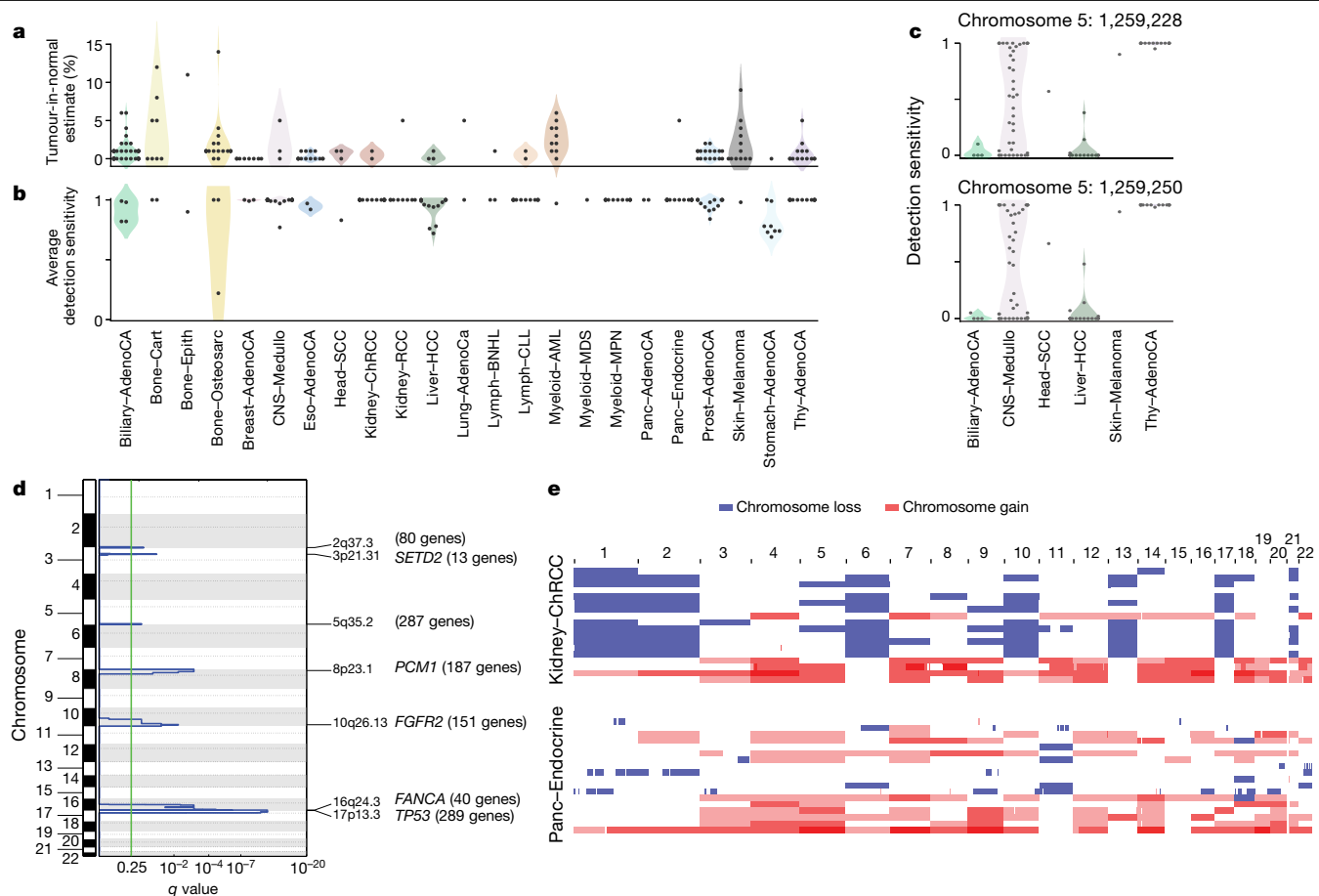
non-coding drivers<sup>34–37</sup> with those newly discovered in PCAWG data; this is reported in a companion paper<sup>4</sup>. Using this approach, only 13% (785 out of 5,913) of driver point mutations were non-coding in PCAWG. Nonetheless, 25% of PCAWG tumours bear at least one putative non-coding driver point mutation, and one third (237 out of 785) affected the *TERT* promoter (9% of PCAWG tumours). Overall, non-coding driver point mutations are less frequent than coding driver mutations. With the exception of the *TERT* promoter, individual enhancers and promoters are only infrequent targets of driver mutations<sup>4</sup>.

Across tumour types, SVs and point mutations have different relative contributions to tumorigenesis. Driver SVs are more prevalent in breast adenocarcinomas ( $6.4 \pm 3.7$  SVs (mean  $\pm$  s.d.) compared with  $2.2 \pm 1.3$  point mutations;  $P < 1 \times 10^{-16}$ , Mann–Whitney  $U$ -test) and ovary adenocarcinomas ( $5.8 \pm 2.6$  SVs compared with  $1.9 \pm 1.0$  point mutations;  $P < 1 \times 10^{-16}$ ), whereas driver point mutations have

patients. **b**, Genomic elements targeted by different types of mutations in the cohort altered in more than 65 tumours. Both germline and somatic variants are included. Left, the heatmap shows the recurrence of alterations across cancer types. The colour indicates the proportion of mutated tumours and the number indicates the absolute count of mutated tumours. Right, the proportion of each type of alteration that affects each genomic element. **c**, Tumour-suppressor genes with biallelic inactivation in 10 or more patients. The values included under the gene labels represent the proportions of patients who have biallelic mutations in the gene out of all patients with a somatic mutation in that gene. GR, genomic rearrangement; SCNA, somatic copy-number alteration; SGR, somatic genome rearrangement; TSG, tumour suppressor gene; UTR, untranslated region.

a larger contribution in colorectal adenocarcinomas ( $2.4 \pm 1.4$  SVs compared with  $7.4 \pm 7.0$  point mutations;  $P = 4 \times 10^{-10}$ ) and mature B cell lymphomas ( $2.2 \pm 1.3$  SVs compared with  $6 \pm 3.8$  point mutations;  $P < 1 \times 10^{-16}$ ), as previously shown<sup>38</sup>. Across tumour types, there are differences in which classes of mutation affect a given genomic element (Fig. 2b).

We confirmed that many driver mutations that affect tumour-suppressor genes are two-hit inactivation events (Fig. 2c). For example, of the 954 tumours in the cohort with driver mutations in *TP53*, 736 (77%) had both alleles mutated, 96% of which (707 out of 736) combined a somatic point mutation that affected one allele with somatic deletion of the other allele. Overall, 17% of patients had rare germline protein-truncating variants (PTVs) in cancer-predisposition genes<sup>39</sup>, DNA-damage response genes<sup>40</sup> and somatic driver genes. Biallelic inactivation due to somatic alteration on top of a germline PTV was observed in 4.5% of patients overall, with 81% of



**Fig. 3 | Analysis of patients with no detected driver mutations. a**, Individual estimates of the percentage of tumour-in-normal contamination across patients with no driver mutations in PCAWG ( $n = 181$ ). No data were available for myelodysplastic syndromes and acute myeloid leukaemia. Points represent estimates for individual patients, and the coloured areas are estimated density distributions (violin plots). Abbreviations of the tumour types are defined in Extended Data Table 1. **b**, Average detection sensitivity by tumour type for tumours without known drivers ( $n = 181$ ). Each dot represents a given sample and is the average sensitivity of detecting clonal substitutions across the genome, taking into account purity and ploidy. Coloured areas are estimated density distributions, shown for cohorts with at least five cases. **c**, Detection

sensitivity for *TERT* promoter hotspots in tumour types in which *TERT* is frequently mutated. Coloured areas are estimated density distributions. **d**, Significant copy-number losses identified by two-sided hypothesis testing using GISTIC2.0, corrected for multiple-hypothesis testing. Numbers in parentheses indicate the number of genes in significant regions when analysing medulloblastomas without known drivers ( $n = 42$ ). Significant regions with known cancer-associated genes are labelled with the representative cancer-associated gene. **e**, Aneuploidy in chromophobe renal cell carcinomas and pancreatic neuroendocrine tumours without known drivers. Patients are ordered on the y axis by tumour type and then by presence of whole-genome duplication (bottom) or not (top).

these affecting known cancer-predisposition genes (such as *BRCA1*, *BRCA2* and *ATM*).

### PCAWG tumours with no apparent drivers

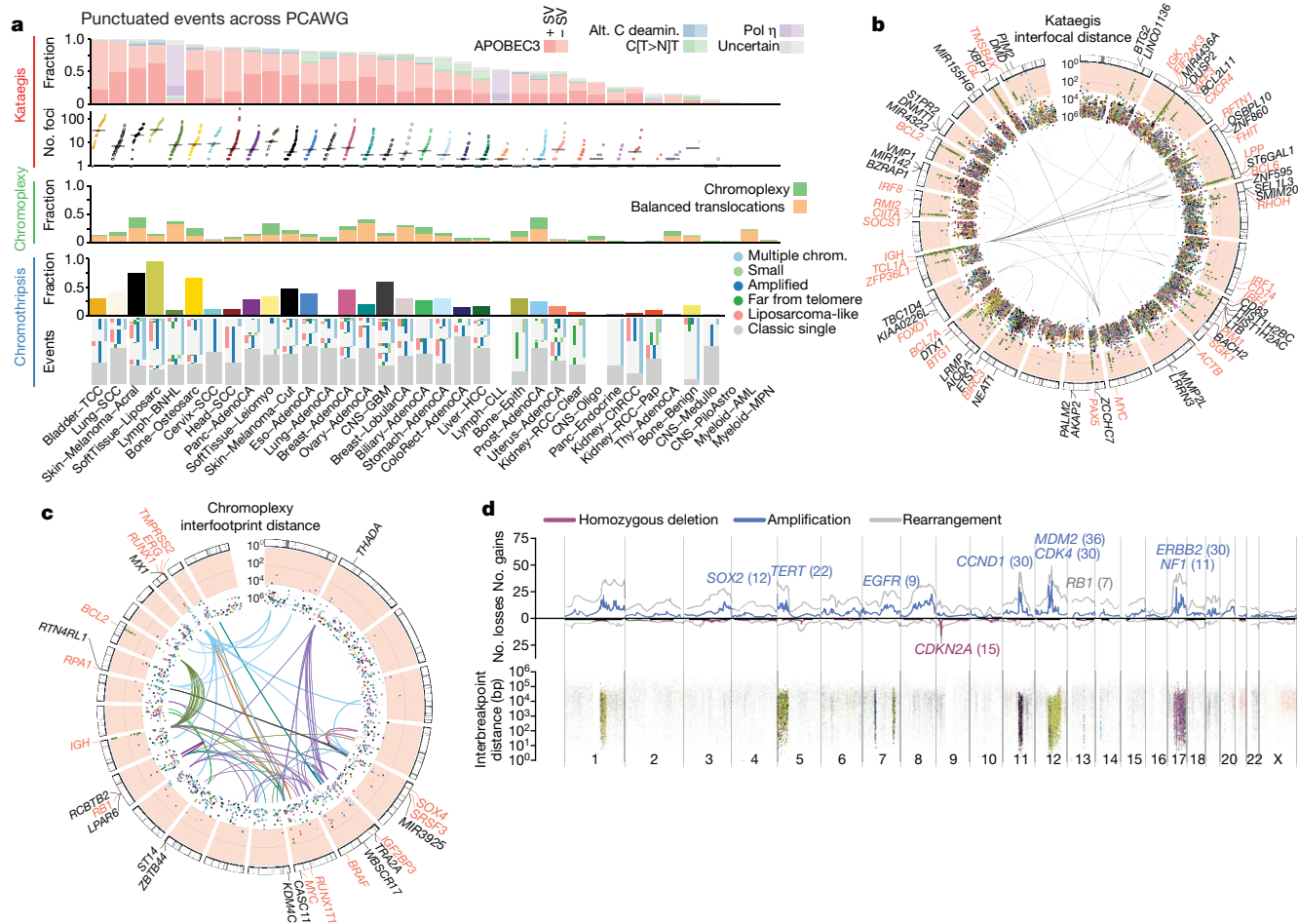
Although more than 90% of PCAWG cases had identified drivers, we found none in 181 tumours (Extended Data Fig. 4a). Reasons for missing drivers have not yet been systematically evaluated in a pan-cancer cohort, and could arise from either technical or biological causes.

Technical explanations could include poor-quality samples, inadequate sequencing or failures in the bioinformatic algorithms used. We assessed the quality of the samples and found that 4 of the 181 cases with no known drivers had more than 5% tumour DNA contamination in their matched normal sample (Fig. 3a). Using an algorithm designed to correct for this contamination<sup>41</sup>, we identified previously missed mutations in genes relevant to the respective cancer types. Similarly, if the fraction of tumour cells in the cancer sample is low through stromal contamination, the detection of driver mutations can be impaired. Most tumours with no known drivers had an average power to detect mutations close to 100%; however, a few had power in the 70–90% range (Fig. 3b and Extended Data Fig. 4b). Even

in adequately sequenced genomes, lack of read depth at specific driver loci can impair mutation detection. For example, only around 50% of PCAWG tumours had sufficient coverage to call a mutation ( $\geq 90\%$  power) at the two *TERT* promoter hotspots, probably because the high GC content of this region causes biased coverage (Fig. 3c). In fact, 6 hepatocellular carcinomas and 2 biliary cholangiocarcinomas among the 181 cases with no known drivers actually did contain *TERT* mutations, which were discovered after deep targeted sequencing<sup>42</sup>.

Finally, technical reasons for missing driver mutations include failures in the bioinformatic algorithms. This affected 35 myeloproliferative neoplasms in PCAWG, in which the *JAK2*<sup>V617F</sup> driver mutation should have been called. Our somatic variant-calling algorithms rely on ‘panels of normals’, typically from blood samples, to remove recurrent sequencing artefacts. As 2–5% of healthy individuals carry occult haematopoietic clones<sup>43</sup>, recurrent driver mutations in these clones can enter panels of normals.

With regard to biological causes, tumours may be driven by mutations in cancer-associated genes that are not yet described for that tumour type. Using driver discovery algorithms on tumours with no known drivers, no individual genes reached significance for point mutations. However, we identified a recurrent CNA that spanned *SETD2* in



**Fig. 4 | Patterns of clustered mutational processes in PCAWG.** **a**, Kataegis. Top, prevalence of different types of kataegis and their association with SVs ( $\leq 1$  kb from the focus). Bottom, the distribution of the number of foci of kataegis per sample. Chromoplexy. Prevalence of chromoplexy across cancer types, subdivided into balanced translocations and more complex events. Chromothripsis. Top, frequency of chromothripsis across cancer types. Bottom, for each cancer type a column is shown, in which each row is a chromothripsis region represented by five coloured rectangles relating to its categorization. **b**, Circos rainfall plot showing the distances between consecutive kataegis events across PCAWG compared with their genomic position. Lymphoid tumours (khaki, B cell non-Hodgkin's lymphoma; orange, chronic lymphocytic leukaemia) have hypermutation hot spots ( $\geq 3$  foci with distance  $\leq 1$  kb; pale red zone), many of which are near known cancer-associated genes (red annotations) and have associated SVs ( $\leq 10$  kb from the focus; shown as arcs in the centre). **c**, Circos rainfall plot as in **b** that shows the distance versus

medulloblastomas that lacked known drivers (Fig. 3d), indicating that restricting hypothesis testing to missing-driver cases can improve power if undiscovered genes are enriched in such tumours. Inactivation of *SETD2* in medulloblastoma significantly decreased gene expression ( $P = 0.002$ ) (Extended Data Fig. 4c). Notably, *SETD2* mutations occurred exclusively in medulloblastoma group-4 tumours ( $P < 1 \times 10^{-4}$ ). Group-4 medulloblastomas are known for frequent mutations in other chromatin-modifying genes<sup>44</sup>, and our results suggest that *SETD2* loss of function is an additional driver that affects chromatin regulators in this subgroup.

Two tumour types had a surprisingly high fraction of patients with out identified driver mutations: chromophobe renal cell carcinoma (44%; 19 out of 43) and pancreatic neuroendocrine cancers (22%; 18 out of 81) (Extended Data Fig. 4a). A notable feature of the missing-driver cases in both tumour types was a remarkably consistent

the position of consecutive chromoplexy and reciprocal translocation footprints across PCAWG. Lymphoid, prostate and thyroid cancers exhibit recurrent events ( $\geq 2$  footprints with distance  $\leq 10$  kb; pale red zone) that are likely to be driver SVs and are annotated with nearby genes and associated SVs, which are shown as bold and thin arcs for chromoplexy and reciprocal translocations, respectively (colours as in **a**). **d**, Effect of chromothripsis along the genome and involvement of PCAWG driver genes. Top, number of chromothripsis-induced gains or losses (grey) and amplifications (blue) or deletions (red). Within the identified chromothripsis regions, selected recurrently rearranged (light grey), amplified (blue) and homozygously deleted (magenta) driver genes are indicated. Bottom, interbreakpoint distance between all subsequent breakpoints within chromothripsis regions across cancer types, coloured by cancer type. Regions with an average interbreakpoint distance  $< 10$  kb are highlighted. C[T>N]T, kataegis with a pattern of thymine mutations in a CpTpT context.

profile of chromosomal aneuploidy—patterns that have previously been reported<sup>45,46</sup> (Fig. 3e). The absence of other identified driver mutations in these patients raises the possibility that certain combinations of whole-chromosome gains and losses may be sufficient to initiate a cancer in the absence of more-targeted driver events such as point mutations or fusion genes of focal CNAs.

Even after accounting for technical issues and novel drivers, 5.3% of PCAWG tumours still had no identifiable driver events. In a research setting, in which we are interested in drawing conclusions about populations of patients, the consequences of technical issues that affect occasional samples will be mitigated by sample size. In a clinical setting, in which we are interested in the driver mutations in a specific patient, these issues become substantially more important. Careful and critical appraisal of the whole pipeline—including sample acquisition, genome sequencing, mapping, variant calling and driver annotation, as done

here—should be required for laboratories that offer clinical sequencing of cancer genomes.

### Patterns of clustered mutations and SVs

Some somatic mutational processes generate multiple mutations in a single catastrophic event, typically clustered in genomic space, leading to substantial reconfiguration of the genome. Three such processes have previously been described: (1) chromoplexy, in which repair of co-occurring double-stranded DNA breaks—typically on different chromosomes—results in shuffled chains of rearrangements<sup>47,48</sup> (Extended Data Fig. 5a); (2) kataegis, a focal hypermutation process that leads to locally clustered nucleotide substitutions, biased towards a single DNA strand<sup>49–51</sup> (Extended Data Fig. 5b); and (3) chromothripsis, in which tens to hundreds of DNA breaks occur simultaneously, clustered on one or a few chromosomes, with near-random stitching together of the resulting fragments<sup>52–55</sup> (Extended Data Fig. 5c). We characterized the PCAWG genomes for these three processes (Fig. 4).

Chromoplexy events and reciprocal translocations were identified in 467 (17.8%) samples (Fig. 4a, c). Chromoplexy was prominent in prostate adenocarcinoma and lymphoid malignancies, as previously described<sup>47,48</sup>, and—unexpectedly—thyroid adenocarcinoma. Different genomic loci were recurrently rearranged by chromoplexy across the three tumour types, mediated by positive selection for particular fusion genes or enhancer-hijacking events. Of 13 fusion genes or enhancer hijacking events in 48 thyroid adenocarcinomas, at least 4 (31%) were caused by chromoplexy, with a further 4 (31%) part of complexes that contained chromoplexy footprints (Extended Data Fig. 5a). These events generated fusion genes that involved *RET* (two cases) and *NTRK3* (one case)<sup>56</sup>, and the juxtaposition of the oncogene *IGF2BP3* with regulatory elements from highly expressed genes (five cases).

Kataegis events were found in 60.5% of all cancers, with particularly high abundance in lung squamous cell carcinoma, bladder cancer, acral melanoma and sarcomas (Fig. 4a, b). Typically, kataegis comprises C > N mutations in a TpC context, which are probably caused by APOBEC activity<sup>49–51</sup>, although a T > N conversion in a TpT or CpT process (the affected T is highlighted in bold) attributed to error-prone polymerases has recently been described<sup>57</sup>. The APOBEC signature accounted for 81.7% of kataegis events and correlated positively with *APOBEC3B* expression levels, somatic SV burden and age at diagnosis (Supplementary Fig. 5). Furthermore, 5.7% of kataegis events involved the T > N error-prone polymerase signature and 2.3% of events, most notably in sarcomas, showed cytidine deamination in an alternative GpC or CpC context.

Kataegis events were frequently associated with somatic SV breakpoints (Fig. 4a and Supplementary Fig. 6a), as previously described<sup>50,51</sup>. Deletions and complex rearrangements were most-strongly associated with kataegis, whereas tandem duplications and other simple SV classes were only infrequently associated (Supplementary Fig. 6b). Kataegis inducing predominantly T > N mutations in CpTpT context was enriched near deletions, specifically those in the 10–25-kilobase (kb) range (Supplementary Fig. 6c).

Samples with extreme kataegis burden (more than 30 foci) comprise four types of focal hypermutation (Extended Data Fig. 6): (1) off-target somatic hypermutation and foci of T > N at CpTpT, found in B cell non-Hodgkin lymphoma and oesophageal adenocarcinomas, respectively; (2) APOBEC kataegis associated with complex rearrangements, notably found in sarcoma and melanoma; (3) rearrangement-independent APOBEC kataegis on the lagging strand and in early-replicating regions, mainly found in bladder and head and neck cancer; and (4) a mix of the last two types. Kataegis only occasionally led to driver mutations (Supplementary Table 5).

We identified chromothripsis in 587 samples (22.3%), most frequently among sarcoma, glioblastoma, lung squamous cell carcinoma, melanoma and breast adenocarcinoma<sup>18</sup>. Chromothripsis

increased with whole-genome duplications in most cancer types (Extended Data Fig. 7a), as previously shown in medulloblastoma<sup>58</sup>. The most recurrently associated driver was *TP53*<sup>52</sup> (pan-cancer odds ratio = 3.22; pan-cancer  $P = 8.3 \times 10^{-35}$ ;  $q < 0.05$  in breast lobular (odds ratio = 13), colorectal (odds ratio = 25), prostate (odds ratio = 2.6) and hepatocellular (odds ratio = 3.9) cancers; Fisher–Boschloo tests). In two cancer types (osteosarcoma and B cell lymphoma), women had a higher incidence of chromothripsis than men (Extended Data Fig. 7b). In prostate cancer, we observed a higher incidence of chromothripsis in patients with late-onset than early-onset disease<sup>59</sup> (Extended Data Fig. 7c).

Chromothripsis regions coincided with 3.6% of all identified drivers in PCAWG and around 7% of copy-number drivers (Fig. 4d). These proportions are considerably enriched compared to expectation if selection were not acting on these events (Extended Data Fig. 7d). The majority of coinciding driver events were amplifications (58%), followed by homozygous deletions (34%) and SVs within genes or promoter regions (8%). We frequently observed a  $\geq 2$ -fold increase or decrease in expression of amplified or deleted drivers, respectively, when these loci were part of a chromothripsis event, compared with samples without chromothripsis (Extended Data Fig. 7e).

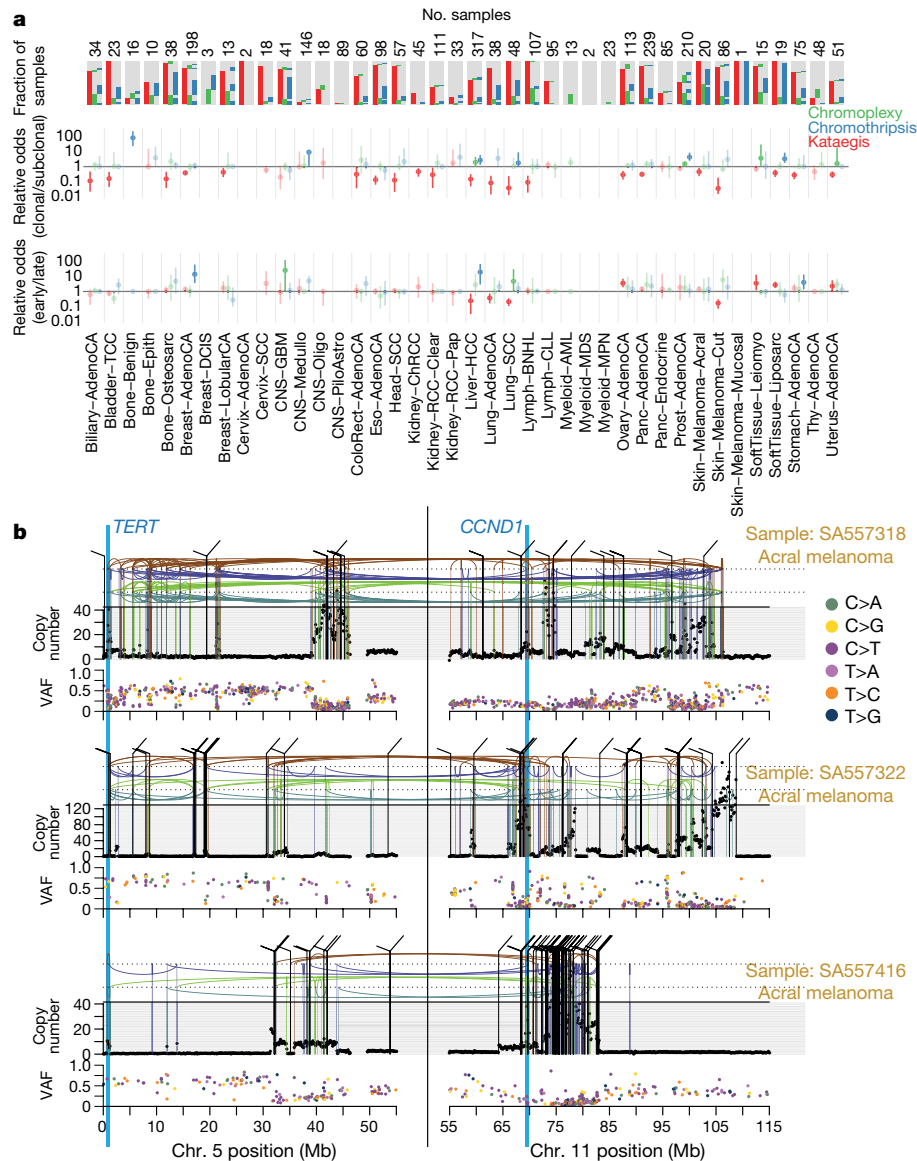
Chromothripsis manifested in diverse patterns and frequencies across tumour types, which we categorized on the basis of five characteristics (Fig. 4a). In liposarcoma, for example, chromothripsis events often involved multiple chromosomes, with universal *MDM2* amplification<sup>60</sup> and co-amplification of *TERT* in 4 of 19 cases (Fig. 4d). By contrast, in glioblastoma the events tended to affect a smaller region on a single chromosome that was distant from the telomere, resulting in focal amplification of *EGFR* and *MDM2* and loss of *CDKN2A*. Acral melanomas frequently exhibited *CCND1* amplification, and lung squamous cell carcinomas *SOX2* amplifications. In both cases, these drivers were more-frequently altered by chromothripsis compared with other drivers in the same cancer type and to other cancer types for the same driver (Fig. 4d and Extended Data Fig. 7f). Finally, in chromophobe renal cell carcinoma, chromothripsis nearly always affected chromosome 5 (Supplementary Fig. 7): these samples had breakpoints immediately adjacent to *TERT*, increasing *TERT* expression by 80-fold on average compared with samples without rearrangements ( $P = 0.0004$ ; Mann–Whitney *U*-test).

### Timing clustered mutations in evolution

An unanswered question for clustered mutational processes is whether they occur early or late in cancer evolution. To address this, we used molecular clocks to define broad epochs in the life history of each tumour<sup>49,61</sup>. One transition point is between clonal and subclonal mutations: clonal mutations occurred before, and subclonal mutations after, the emergence of the most-recent common ancestor. In regions with copy-number gains, molecular time can be further divided according to whether mutations preceded the copy-number gain (and were themselves duplicated) or occurred after the gain (and therefore present on only one chromosomal copy)<sup>7</sup>.

Chromothripsis tended to have greater relative odds of being clonal than subclonal, suggesting that it occurs early in cancer evolution, especially in liposarcomas, prostate adenocarcinoma and squamous cell lung cancer (Fig. 5a). As previously reported, chromothripsis was especially common in melanomas<sup>62</sup>. We identified 89 separate chromothripsis events that affected 66 melanomas (61%); 47 out of 89 events affected genes known to be recurrently altered in melanoma<sup>63</sup> (Supplementary Table 6). Involvement of a region on chromosome 11 that includes the cell-cycle regulator *CCND1* occurred in 21 cases (10 out of 86 cutaneous, and 11 out of 21 acral or mucosal melanomas), typically combining chromothripsis with amplification (19 out of 21 cases) (Extended Data Fig. 8). Co-involvement of other cancer-associated genes in the same chromothripsis event was also frequent, including





**Fig. 5 | Timing of clustered events in PCAWG. a**, Extent and timing of chromothripsis, kataegis and chromoplexy across PCAWG. Top, stacked bar charts illustrate co-occurrence of chromothripsis, kataegis and chromoplexy in the samples. Middle, relative odds of clustered events being clonal or subclonal are shown with bootstrapped 95% confidence intervals. Point estimates are highlighted when they do not overlap odds of 1:1. Bottom, relative odds of the events being early or late clonal are shown as above. Sample

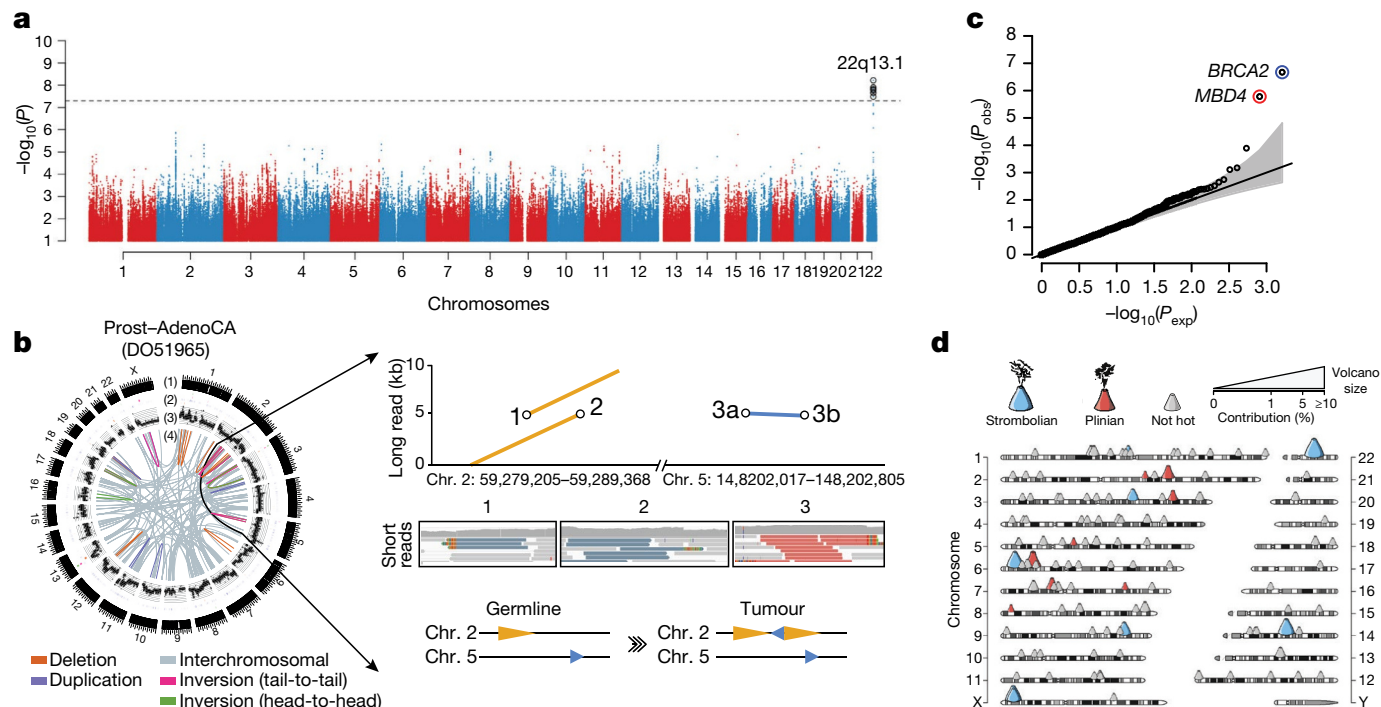
sizes (number of patients) are shown across the top. **b**, Three representative patients with acral melanoma and chromothripsis-induced amplification that simultaneously affects *TERT* and *CCND1*. The black points (top) represent sequence coverage from individual genomic bins, with SVs shown as coloured arcs (translocation in black, deletion in purple, duplication in brown, tail-to-tail inversion in cyan and head-to-head inversion in green). Bottom, the variant allele fractions of somatic point mutations.

*TERT* (five cases), *CDKN2A* (three cases), *TP53* (two cases) and *MYC* (two cases) (Fig. 5b). In these co-amplifications, a chromothripsis event involving multiple chromosomes initiated the process, creating a derivative chromosome in which hundreds of fragments were stitched together in a near-random order (Fig. 5b). This derivative then rearranged further, leading to massive co-amplification of the multiple target oncogenes together with regions located nearby on the derivative chromosome.

In these cases of amplified chromothripsis, we can use the inferred number of copies bearing each SNV to time the amplification process. SNVs present on the chromosome before amplification will themselves be amplified and are therefore reported in a high fraction of sequence reads (Fig. 5b and Extended Data Fig. 8). By contrast, late SNVs that occur after the amplification has concluded will be present on only one chromosome copy out of many, and thus have a low variant

allele fraction. Regions of *CCND1* amplification had few—sometimes zero—mutations at high variant allele fraction in acral melanomas, in contrast to later *CCND1* amplifications in cutaneous melanomas, in which hundreds to thousands of mutations typically predated amplification (Fig. 5b and Extended Data Fig. 9a, b). Thus, both chromothripsis and the subsequent amplification generally occurred very early during the evolution of acral melanoma. By comparison, in lung squamous cell carcinomas, similar patterns of chromothripsis followed by *SOX2* amplification are characterized by many amplified SNVs, suggesting a later event in the evolution of these cancers (Extended Data Fig. 9c).

Notably, in cancer types in which the mutational load was sufficiently high, we could detect a larger-than-expected number of SNVs on an intermediate number of DNA copies, suggesting that they appeared during the amplification process (Supplementary Fig. 8).



**Fig. 6 | Germline determinants of the somatic mutation landscape.**

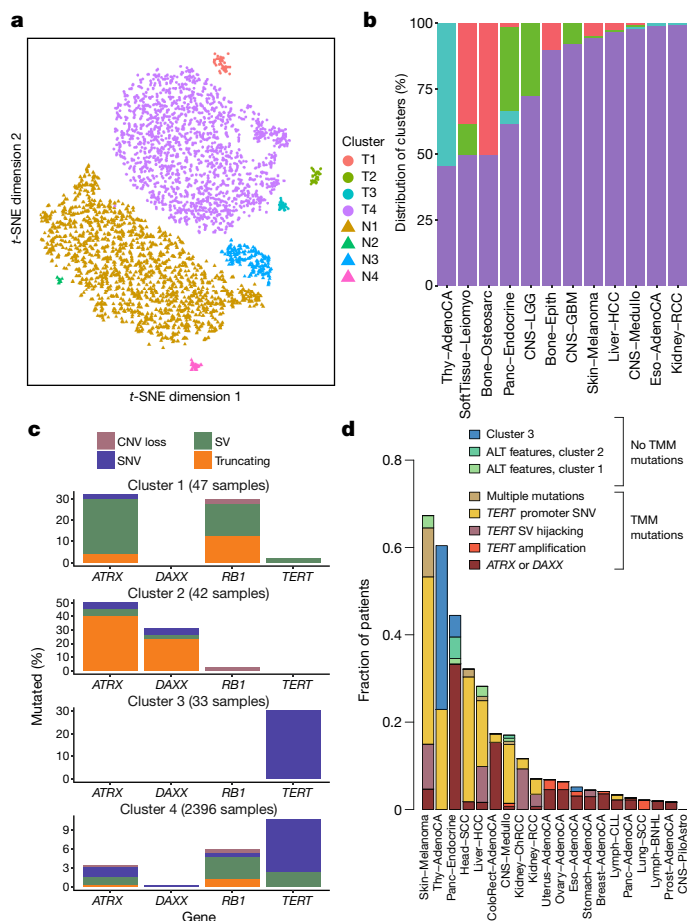
**a**, Association between common (MAF > 5%) germline variants and somatic APOBEC3B-like mutagenesis in individuals of European ancestry ( $n = 1,201$ ). Two-sided hypothesis testing was performed with PLINK v1.9. To mitigate multiple-hypothesis testing, the significance threshold was set to genome-wide significance ( $P < 5 \times 10^{-8}$ ). **b**, Templated insertion SVs in a *BRCA1*-associated prostate cancer. Left, chromosome bands (1); SVs  $\leq 10$  megabases (Mb) (2); 1-kb read depth corrected to copy number 0–6 (3); inter- and intrachromosomal SVs > 10 Mb (4). Right, a complex somatic SV composed of a 2.2-kb tandem duplication on chromosome 2 together with a 232-base-pair (bp) inverted templated insertion SV that is derived from chromosome 5 and inserted inbetween the tandem duplication (bottom). Consensus sequence alignment of locally assembled Oxford Nanopore Technologies long sequencing reads to chromosomes 2 and 5 of the human reference genome (top). Breakpoints are circled and marked as 1 (beginning of tandem duplication), 2 (end of tandem duplication) or 3 (inverted templated insertion). For each breakpoint, the middle panel shows Illumina short reads at SV

breakpoints. **c**, Association between rare germline PTVs (MAF < 0.5%) and somatic CpG mutagenesis (approximately with signature 1) in individuals of European ancestry ( $n = 1,201$ ). Genes highlighted in blue or red were associated with lower or higher somatic mutation rates. Two-sided hypothesis testing was performed using linear-regression models with sex, age at diagnosis and cancer project as variables. To mitigate multiple-hypothesis testing, the significance threshold was set to exome-wide significance ( $P < 2.5 \times 10^{-6}$ ). The black line represents the identity line that would be followed if the observed  $P$  values followed the null expectation; the shaded area shows the 95% confidence intervals. **d**, Catalogue of polymorphic germline L1 source elements that are active in cancer. The chromosomal map shows germline source L1 elements as volcano symbols. Each volcano is colour-coded according to the type of source L1 activity. The contribution of each source locus (expressed as a percentage) to the total number of transductions identified in PCAWG tumours is represented as a gradient of volcano size, with top contributing elements exhibiting larger sizes.

## Germline effects on somatic mutations

We integrated the set of 88 million germline genetic variant calls with somatic mutations in PCAWG, to study germline determinants of somatic mutation rates and patterns. First, we performed a genome-wide association study of somatic mutational processes with common germline variants (minor allele frequency (MAF) > 5%) in individuals with inferred European ancestry. An independent genome-wide association study was performed in East Asian individuals from Asian cancer genome projects. We focused on two prevalent endogenous mutational processes: spontaneous deamination of 5-methylcytosine at CpG dinucleotides<sup>5</sup> (signature 1) and activity of the APOBEC3 family of cytidine deaminases<sup>64</sup> (signatures 2 and 13). No locus reached genome-wide significance ( $P < 5 \times 10^{-8}$ ) for signature 1 (Extended Data Fig. 10a, b). However, a locus at 22q13.1 predicted an APOBEC3B-like mutagenesis at the pan-cancer level<sup>65</sup> (Fig. 6a). The strongest signal at 22q13.1 was driven by rs12628403, and the minor (non-reference) allele was protective against APOBEC3B-like mutagenesis ( $\beta = -0.43$ ,  $P = 5.6 \times 10^{-9}$ , MAF = 8.2%,  $n = 1,201$  donors) (Extended Data Fig. 10c). This variant tags a common, approximately 30-kb germline SV that deletes the *APOBEC3B* coding sequence and fuses the *APOBEC3B* 3' untranslated region with the coding sequence of *APOBEC3A*. The deletion is known

to increase breast cancer risk and APOBEC3 mutagenesis in breast cancer genomes<sup>66,67</sup>. Here, we found that rs12628403 reduces APOBEC3B-like mutagenesis specifically in cancer types with low levels of APOBEC3 mutagenesis ( $\beta_{\text{low}} = -0.50$ ,  $P_{\text{low}} = 1 \times 10^{-8}$ ;  $\beta_{\text{high}} = 0.17$ ,  $P_{\text{high}} = 0.2$ ), and increases APOBEC3A-like mutagenesis in cancer types with high levels of APOBEC3 mutagenesis ( $\beta_{\text{high}} = 0.44$ ,  $P_{\text{high}} = 8 \times 10^{-4}$ ;  $\beta_{\text{low}} = -0.21$ ,  $P_{\text{low}} = 0.02$ ). Moreover, we identified a second, novel locus at 22q13.1 that was associated with APOBEC3B-like mutagenesis across cancer types (rs2142833,  $\beta = 0.23$ ,  $P = 1.3 \times 10^{-8}$ ). We independently validated the association between both loci and APOBEC3B-like mutagenesis using East Asian individuals from Asian cancer genome projects ( $\beta_{\text{rs12628403}} = 0.57$ ,  $P_{\text{rs12628403}} = 4.2 \times 10^{-12}$ ;  $\beta_{\text{rs2142833}} = 0.58$ ,  $P_{\text{rs2142833}} = 8 \times 10^{-15}$ ) (Extended Data Fig. 10d). Notably, in a conditional analysis that accounted for rs12628403, we found that rs2142833 and rs12628403 are inherited independently in Europeans ( $r^2 < 0.1$ ), and rs2142833 remained significantly associated with APOBEC3B-like mutagenesis in Europeans ( $\beta_{\text{EUR}} = 0.17$ ,  $P_{\text{EUR}} = 3 \times 10^{-5}$ ) and East Asians ( $\beta_{\text{ASN}} = 0.25$ ,  $P_{\text{ASN}} = 2 \times 10^{-3}$ ) (Extended Data Fig. 10e, f). Analysis of donor-matched expression data further suggests that rs2142833 is a *cis*-expression quantitative trait locus (eQTL) for *APOBEC3B* at the pan-cancer level ( $\beta = 0.19$ ,  $P = 2 \times 10^{-6}$ ) (Extended Data Fig. 10g, h), consistent with *cis*-eQTL studies in normal cells<sup>68,69</sup>.



**Fig. 7 | Telomere sequence patterns across PCAWG.** **a**, Scatter plot of the clusters of telomere patterns identified across PCAWG using *t*-distributed stochastic neighbour embedding (*t*-SNE), based on  $n = 2,518$  tumour samples and their matched normal samples. Axes have arbitrary dimensions such that samples with similar telomere profiles are clustered together and samples with dissimilar telomere profiles are far apart with high probability. **b**, Distribution of the four tumour-specific clusters of telomere patterns in selected tumour types from PCAWG. **c**, Distribution of relevant driver mutations associated with alternative lengthening of telomere and normal telomere maintenance across the four clusters. **d**, Distribution of telomere maintenance abnormalities across tumour types with more than 40 patients in PCAWG. Samples were classified as tumour clusters 1–3 if they fell into a relevant cluster without mutations in *TERT*, *ATRX* or *DAXX* and had no ALT phenotype. TMM, telomere maintenance mechanisms.

Second, we performed a rare-variant association study ( $MAF < 0.5\%$ ) to investigate the relationship between germline PTVs and somatic DNA rearrangements in individuals with European ancestry (Extended Data Fig. 11a–c). Germline *BRCA2* and *BRCA1* PTVs were associated with an increased burden of small (less than 10 kb) somatic SV deletions ( $P = 1 \times 10^{-8}$ ) and tandem duplications ( $P = 6 \times 10^{-13}$ ), respectively, corroborating recent studies in breast and ovarian cancer<sup>30,70</sup>. In PCAWG data, this pattern also extends to other tumour types, including adenocarcinomas of the prostate and pancreas<sup>6</sup>, typically in the setting of biallelic inactivation. In addition, tumours with high levels of small SV tandem duplications frequently exhibited a novel and distinct class of SVs termed ‘cycles of templated insertions’<sup>6</sup>. These complex SV events consist of DNA templates that are copied from across the genome, joined into one contiguous sequence and inserted into a single derivative chromosome. We found a significant association between germline *BRCA1* PTVs and templated insertions at the pan-cancer level ( $P = 4 \times 10^{-15}$ ) (Extended Data Fig. 11d,e). Whole-genome

long-read sequencing data generated for a *BRCA1*-deficient PCAWG prostate tumour verified the small tandem-duplication and templated-insertion SV phenotypes (Fig. 6b). Almost all (20 out of 21) of *BRCA1*-associated tumours with a templated-insertion SV phenotype displayed combined germline and somatic hits in the gene. Together, these data suggest that biallelic inactivation of *BRCA1* is a driver of the templated-insertion SV phenotype.

Third, rare-variant association analysis revealed that patients with germline *MBD4* PTVs had increased rates of somatic C > T mutation rates at CpG dinucleotides ( $P < 2.5 \times 10^{-6}$ ) (Fig. 6c and Extended Data Fig. 11f,g). Analysis of previously published whole-exome sequencing samples from the TCGA ( $n = 8,134$ ) replicated the association between germline *MBD4* PTVs and increased somatic CpG mutagenesis at the pan-cancer level ( $P = 7.1 \times 10^{-4}$ ) (Extended Data Fig. 11h). Moreover, gene-expression profiling revealed a significant but modest correlation between *MBD4* expression and somatic CpG mutation rates between and within PCAWG tumour types (Extended Data Fig. 11i–k). *MBD4* encodes a DNA-repair gene that removes thymidines from T:G mismatches within methylated CpG sites<sup>71</sup>, a functionality that would be consistent with a CpG mutational signature in cancer.

Fourth, we assessed long interspersed nuclear elements (LINE-1; L1 hereafter) that mediate somatic retrotransposition events<sup>72–74</sup>. We identified 114 germline source L1 elements capable of active somatic retrotransposition, including 70 that represent insertions with respect to the human reference genome (Fig. 6d and Supplementary Table 7), and 53 that were tagged by single-nucleotide polymorphisms in strong linkage disequilibrium (Supplementary Table 7). Only 16 germline L1 elements accounted for 67% (2,440 out of 3,669) of all L1-mediated transductions<sup>10</sup> detected in the PCAWG dataset (Extended Data Fig. 12a). These 16 hot-L1 elements followed two broad patterns of somatic activity (8 of each), which we term Strombolian and Plinian in analogy to patterns of volcanic activity. Strombolian L1s are frequently active in cancer, but mediate only small-to-modest eruptions of somatic L1 activity in cancer samples (Extended Data Fig. 12b). By contrast, Plinian L1s are more rarely seen, but display aggressive somatic activity. Whereas Strombolian elements are typically relatively common ( $MAF > 2\%$ ) and sometimes even fixed in the human population, all Plinian elements were infrequent ( $MAF \leq 2\%$ ) in PCAWG donors (Extended Data Fig. 12c;  $P = 0.001$ , Mann–Whitney *U*-test). This dichotomous pattern of activity and allele frequency may reflect differences in age and selective pressures, with Plinian elements potentially inserted into the human germline more recently. PCAWG donors bear on average between 50 and 60 L1 source elements and between 5 and 7 elements with hot activity (Extended Data Fig. 12d), but only 38% (1,075 out of 2,814) of PCAWG donors carried  $\geq 1$  Plinian element. Some L1 germline source loci caused somatic loss of tumour-suppressor genes (Extended Data Fig. 12e). Many are restricted to individual continental population ancestries (Extended Data Fig. 12f–j).

## Replicative immortality

One of the hallmarks of cancer is the ability of cancer to evade cellular senescence<sup>21</sup>. Normal somatic cells typically have finite cell division potential; telomere attrition is one mechanism to limit numbers of mitoses<sup>75</sup>. Cancers enlist multiple strategies to achieve replicative immortality. Overexpression of the telomerase gene, *TERT*, which maintains telomere lengths, is especially prevalent. This can be achieved through point mutations in the promoter that lead to de novo transcription factor binding<sup>34,37</sup>; hitching *TERT* to highly active regulatory elements elsewhere in the genome<sup>46,76</sup>; insertions of viral enhancers upstream of the gene<sup>77,78</sup>; and increased dosage through chromosomal amplification, as we have seen in melanoma (Fig. 5b). In addition, there is an ‘alternative lengthening of telomeres’ (ALT) pathway, in which telomeres are lengthened through homologous recombination, mediated by loss-of-function mutations in the *ATRX* and *DAXX* genes<sup>79</sup>.

As reported in a companion paper<sup>13</sup>, 16% of tumours in the PCAWG dataset exhibited somatic mutations in at least one of *ATRX*, *DAXX* and *TERT*. *TERT* alterations were detected in 270 samples, whereas 128 tumours had alterations in *ATRX* or *DAXX*, of which 71 were protein-truncating. In the companion paper, which focused on describing patterns of ALT and *TERT*-mediated telomere maintenance<sup>13</sup>, 12 features of telomeric sequence were measured in the PCAWG cohort. These included counts of nine variants of the core hexameric sequence, the number of ectopic telomere-like insertions within the genome, the number of genomic breakpoints and telomere length as a ratio between tumour and normal. Here we used the 12 features as an overview of telomere integrity across all tumours in the PCAWG dataset.

On the basis of these 12 features, tumour samples formed 4 distinct subclusters (Fig. 7a and Extended Data Fig. 13a), suggesting that telomere-maintenance mechanisms are more diverse than the well-established *TERT* and ALT dichotomy. Clusters C1 (47 tumours) and C2 (42 tumours) were enriched for traits of the ALT pathway—having longer telomeres, more genomic breakpoints, more ectopic telomere insertions and variant telomere sequence motifs (Supplementary Fig. 9). C1 and C2 were distinguished from one another by the latter having a considerable increase in the number of TTCGGG and TGAGGG variant motifs among the telomeric hexamers. Thyroid adenocarcinomas were markedly enriched among C3 samples (26 out of 33 C3 samples;  $P < 10^{-16}$ ); the C1 cluster (ALT subtype 1) was common among sarcomas; and both pancreatic endocrine neoplasms and low-grade gliomas had a high proportion of samples in the C2 cluster (ALT subtype 2) (Fig. 7b). Notably, some of the thyroid adenocarcinomas and pancreatic neuroendocrine tumours that cluster together (cluster C3) had matched normal samples that also cluster together (normal cluster N3) (Extended Data Fig. 13a) and which share common properties. For example, the GTAGGG repeat was overrepresented among samples in this group (Supplementary Fig. 10).

Somatic driver mutations were also unevenly distributed across the four clusters (Fig. 7c). C1 tumours were enriched for *RBI* mutations or SVs ( $P = 3 \times 10^{-5}$ ), as well as frequent SVs that affected *ATRX* ( $P = 6 \times 10^{-14}$ ), but not *DAXX*. *RBI* and *ATRX* mutations were largely mutually exclusive (Extended Data Fig. 13b). By contrast, C2 tumours were enriched for somatic point mutations in *ATRX* and *DAXX* ( $P = 6 \times 10^{-5}$ ), but not *RBI*. The enrichment of *RBI* mutations in C1 remained significant when only leiomyosarcomas and osteosarcomas were considered, confirming that this enrichment is not merely a consequence of the different distribution of tumour types across clusters. C3 samples had frequent *TERT* promoter mutations (30%;  $P = 2 \times 10^{-6}$ ).

There was a marked predominance of *RBI* mutations in C1. Nearly a third of the samples in C1 contained an *RBI* alteration, which were evenly distributed across truncating SNVs, SVs and shallow deletions (Extended Data Fig. 13c). Previous research has shown that *RBI* mutations are associated with long telomeres in the absence of *TERT* mutations and *ATRX* inactivation<sup>80</sup>, and studies using mouse models have shown that knockout of Rb-family proteins causes elongated telomeres<sup>81</sup>. The association with the C1 cluster here suggests that *RBI* mutations can represent another route to activating the ALT pathway, which has subtly different properties of telomeric sequence compared with the inactivation of *DAXX*—these fall almost exclusively in cluster C2.

Tumour types with the highest rates of abnormal telomere maintenance mechanisms often originate in tissues that have low endogenous replicative activity (Fig. 7d). In support of this, we found an inverse correlation between previously estimated rates of stem cell division across tissues<sup>82</sup> and the frequency of telomere maintenance abnormalities ( $P = 0.01$ , Poisson regression) (Extended Data Fig. 13d). This suggests that restriction of telomere maintenance is an important tumour-suppression mechanism, particularly in tissues with low steady-state cellular proliferation, in which a clone must overcome this constraint to achieve replicative immortality.

## Conclusions and future perspectives

The resource reported in this paper and its companion papers has yielded insights into the nature and timing of the many mutational processes that shape large- and small-scale somatic variation in the cancer genome; the patterns of selection that act on these variations; the widespread effect of somatic variants on transcription; the complementary roles of the coding and non-coding genome for both germline and somatic mutations; the ubiquity of intratumoral heterogeneity; and the distinctive evolutionary trajectory of each cancer type. Many of these insights can be obtained only from an integrated analysis of all classes of somatic mutation on a whole-genome scale, and would not be accessible with, for example, targeted exome sequencing.

The promise of precision medicine is to match patients to targeted therapies using genomics. A major barrier to its evidence-based implementation is the daunting heterogeneity of cancer chronicled in these papers, from tumour type to tumour type, from patient to patient, from clone to clone and from cell to cell. Building meaningful clinical predictors from genomic data can be achieved, but will require knowledge banks comprising tens of thousands of patients with comprehensive clinical characterization<sup>83</sup>. As these sample sizes will be too large for any single funding agency, pharmaceutical company or health system, international collaboration and data sharing will be required. The next phase of ICGC, ICGC-ARGO (<https://www.icgc-argo.org/>), will bring the cancer genomics community together with healthcare providers, pharmaceutical companies, data science and clinical trials groups to build comprehensive knowledge banks of clinical outcome and treatment data from patients with a wide variety of cancers, matched with detailed molecular profiling.

Extending the story begun by TCGA, ICGC and other cancer genomics projects, the PCAWG has brought us closer to a comprehensive narrative of the causal biological changes that drive cancer phenotypes. We must now translate this knowledge into sustainable, meaningful clinical treatments.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1969-6>.

1. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
2. Pleasance, E. D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
3. Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
4. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,693 cancer whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1965-x> (2020).
5. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* <https://doi.org/10.1038/s41586-020-1943-3> (2020).
6. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* <https://doi.org/10.1038/s41586-019-1913-9> (2020).
7. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* <https://doi.org/10.1038/s41586-019-1907-7> (2020).
8. PCAWG Transcriptome Core Group et al. Genomic basis of RNA alterations in cancer. *Nature* <https://doi.org/10.1038/s41586-020-1970-0> (2020).
9. Zhang, Y. et al. High-coverage whole-genome analysis of 1,220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nat. Commun.* <https://doi.org/10.1038/s41467-019-13885-w> (2020).
10. Rodríguez-Martín, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0562-0> (2020).
11. Zappatka, M. et al. The landscape of viral associations in human cancers. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0558-9> (2020).
12. Jiao, W. et al. A deep learning system can accurately classify primary and metastatic cancers based on patterns of passenger mutations. *Nat. Commun.* <https://doi.org/10.1038/s41467-019-13825-8> (2020).



13. Sieverling, L. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun.* <https://doi.org/10.1038/s41467-019-13824-9> (2020).
14. Yuan, Y. et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0557-x> (2020).
15. Akdemir, K. C. et al. Chromatin folding domains disruptions by somatic genomic rearrangements in human cancers. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0564-y> (2020).
16. Reyna, M. A. et al. Pathway and network analysis of more than 2,500 whole cancer genomes. *Nat. Commun.* <https://doi.org/10.1038/s41467-020-14351-8> (2020).
17. Bailey, M. H. et al. Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. *Nat. Commun.* (2020).
18. Cortes-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0576-7> (2020).
19. Bray, F., Ren, J.-S., Masuyer, E. & Ferlay, J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer* **132**, 1133–1145 (2013).
20. Tarver, T. Cancer Facts & Figures 2012. American Cancer Society (ACS). *J. Consum. Health Internet* **16**, 366–367 (2012).
21. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
22. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
23. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
24. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
25. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
26. Stein, L. D., Knoppers, B. M., Campbell, P., Getz, G. & Korbel, J. O. Data analysis: create a cloud commons. *Nature* **523**, 149–151 (2015).
27. Phillips, M. et al. Genomics: data sharing needs international code of conduct. *Nature* <https://doi.org/10.1038/d41586-020-00082-9> (2020).
28. Krochmalski, J. *Developing with Docker* (Packt Publishing, 2016).
29. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
30. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
31. Meier, B. et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.* **24**, 1624–1636 (2014).
32. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
33. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
34. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
35. Rheinbay, E. et al. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).
36. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
37. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
38. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
39. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).
40. Pearl, L. H., Schierz, A. C., Ward, S. E., Al-Lazikani, B. & Pearl, F. M. G. Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer* **15**, 166–180 (2015).
41. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).
42. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
43. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
44. Northcott, P. A. et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
45. Scarpa, A. et al. Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature* **543**, 65–71 (2017).
46. Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
47. Berger, M. F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
48. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
49. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
50. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
51. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
52. Rausch, T. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71 (2012).
53. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
54. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
55. Zhang, C.-Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
56. The Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
57. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
58. Mardin, B. R. et al. A cell-based model system links chromothripsis with hyperploidy. *Mol. Syst. Biol.* **11**, 828 (2015).
59. Weischenfeldt, J. et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159–170 (2013).
60. Garsed, D. W. et al. The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**, 653–667 (2014).
61. Durinck, S. et al. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* **1**, 137–143 (2011).
62. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
63. The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
64. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
65. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072 (2015).
66. Nik-Zainal, S. et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **46**, 487–491 (2014).
67. Middlebrooks, C. D. et al. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* **48**, 1330–1338 (2016).
68. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
69. Stranger, B. E. et al. Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
70. Menghi, F. et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc. Natl. Acad. Sci. USA* **113**, E2373–E2382 (2016).
71. Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301–304 (1999).
72. Lee, E. et al. Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
73. Tubio, J. M. C. et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343–1251343 (2014).
74. Helman, E. et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).
75. Shay, J. W. & Wright, W. E. Hayflick, his limit, and cellular ageing. *Nat. Rev. Mol. Cell Biol.* **1**, 72–76 (2000).
76. Peifer, M. et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).
77. Totoki, Y. et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat. Genet.* **46**, 1267–1273 (2014).
78. Paterlini-Bréchet, P. et al. Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. *Oncogene* **22**, 3911–3916 (2003).
79. Heaphy, C. M. et al. Prevalence of the alternative lengthening of telomeres telomere maintenance mechanism in human cancer subtypes. *Am. J. Pathol.* **179**, 1608–1615 (2011).
80. Barthel, F. P. et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
81. García-Cao, M., Gonzalo, S., Dean, D. & Blasco, M. A. A role for the Rb family of proteins in controlling telomere length. *Nat. Genet.* **32**, 415–419 (2002).
82. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
83. Gerstung, M. et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49**, 332–340 (2017).
84. O'Connor, B. D. et al. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000Res.* **6**, 52 (2017).
85. Zhang, J. et al. The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
86. Miller, C. A., Qiao, Y., DiSera, T., D'Astous, B. & Marth, G. T. bam.io: a web-based, real-time, sequence alignment file inspector. *Nat. Methods* **11**, 1189–1189 (2014).
87. Goldman, M. et al. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. Preprint at <https://www.biorxiv.org/content/10.1101/326470v6> (2019).
88. Papatheodorou, I. et al. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* **46**, D246–D251 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

# Article

## The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Peter J. Campbell<sup>1,2,745\*</sup>, Gad Getz<sup>3,4,5,6,745\*</sup>, Jan O. Korbe<sup>7,8,745\*</sup>, Joshua M. Stuart<sup>8,745\*</sup>, Jennifer L. Jennings<sup>10,11,745</sup>, Lincoln D. Stein<sup>12,13,745\*</sup>, Marc D. Perry<sup>14,15</sup>, Hardeep K. Nahal-Bose<sup>15</sup>, B. F. Francis Ouellette<sup>16,17</sup>, Constance H. Li<sup>12,18</sup>, Esther Rheinbay<sup>3,6,19</sup>, G. Petur Nielsen<sup>19</sup>, Dennis C. Sgroi<sup>19</sup>, Chin-Lee Wu<sup>19</sup>, William C. Faquin<sup>19</sup>, Vikram Deshpande<sup>19</sup>, Paul C. Boutros<sup>12,18,20,21</sup>, Alexander J. Lazar<sup>22</sup>, Katherine A. Hoadley<sup>23,24</sup>, David N. Louis<sup>19</sup>, L. Jonathan Dursi<sup>12,25</sup>, Christina K. Yung<sup>15</sup>, Matthew H. Bailey<sup>26,27</sup>, Gordon Saksena<sup>3</sup>, Keiran M. Raine<sup>1</sup>, Ivo Buchhalter<sup>28,29,30</sup>, Kortine Kleinheinz<sup>28,30</sup>, Matthias Schlesner<sup>28,31</sup>, Junjun Zhang<sup>15</sup>, Wenyi Wang<sup>32</sup>, David A. Wheeler<sup>33,34</sup>, Li Ding<sup>26,27,35</sup>, Jared T. Simpson<sup>12,36</sup>, Brian D. O'Connor<sup>15,37</sup>, Sergei Yakneen<sup>8</sup>, Kyle Ellrott<sup>38</sup>, Naoki Miyoshi<sup>39</sup>, Adam P. Butler<sup>1</sup>, Romina Royo<sup>40</sup>, Solomon I. Shorser<sup>12</sup>, Miguel Vazquez<sup>40,41</sup>, Tobias Rausch<sup>8</sup>, Grace Tiao<sup>3</sup>, Sebastian M. Waszak<sup>8</sup>, Bernardo Rodriguez-Martin<sup>42,43,44</sup>, Suyash Shringarpure<sup>45</sup>, Dai-Ying Wu<sup>46</sup>, German M. Demidov<sup>47,48,49</sup>, Olivier Delaneau<sup>50,51,52</sup>, Shuto Hayashi<sup>39</sup>, Seiya Imoto<sup>39</sup>, Nina Habermann<sup>8</sup>, Ayellet V. Segre<sup>3,53</sup>, Erik Garrison<sup>1</sup>, Andy Cafferkey<sup>7</sup>, Eva G. Alvarez<sup>42,43,44</sup>, José María Heredia-Genestar<sup>54</sup>, Francesc Muyas<sup>47,48,49</sup>, Oliver Drechsel<sup>47,49</sup>, Alicia L. Bruzos<sup>42,43,44</sup>, Javier Temes<sup>42,43</sup>, Jorge Zamora<sup>1,42,43,44</sup>, Adrian Baez-Ortega<sup>55</sup>, Hyung-Lae Kim<sup>56</sup>, R. Jay Mashl<sup>27,57</sup>, Kai Ye<sup>58,59</sup>, Anthony DiBiase<sup>60</sup>, Kuan-lin Huang<sup>27,61</sup>, Ivica Letunic<sup>62</sup>, Michael D. McLellan<sup>26,27,35</sup>, Steven J. Newhouse<sup>7</sup>, Tal Shmaya<sup>46</sup>, Shushant Kumar<sup>63,64</sup>, David C. Wedge<sup>165,66</sup>, Mark H. Wright<sup>45</sup>, Venkata D. Yellapantula<sup>67,68</sup>, Mark Gerstein<sup>63,64,69</sup>, Ekta Khurana<sup>70,71,72,73</sup>, Tomas Marques-Bonet<sup>74,75,76,77</sup>, Arcadi Navarro<sup>74,75,76</sup>, Carlos D. Bustamante<sup>78</sup>, Reiner Siebert<sup>79,80</sup>, Hidewaki Nakagawa<sup>81</sup>, Douglas F. Easton<sup>82,83</sup>, Stephan Ossowski<sup>47,48,49</sup>, Jose M. C. Tubio<sup>42,43,44</sup>, Francisco M. De La Vega<sup>45,46,78</sup>, Xavier Estivill<sup>47,84</sup>, Denis Yuen<sup>12</sup>, George L. Mihaiescu<sup>15</sup>, Larsson Omberg<sup>85</sup>, Vincent Ferretti<sup>15,86</sup>, Radhakrishnan Sabarinathan<sup>87,88,89</sup>, Oriol Pich<sup>87,89</sup>, Abel Gonzalez-Perez<sup>87,89</sup>, Amaro Taylor-Weiner<sup>90</sup>, Matthew W. Fittall<sup>91</sup>, Jonas Demeulemeester<sup>91,92</sup>, Maxime Tarabichi<sup>191</sup>, Nicola D. Roberts<sup>1</sup>, Peter Van Lo<sup>91,92</sup>, Isidro Cortes-Ciriano<sup>93,94,95</sup>, Lara Urban<sup>78</sup>, Peter Park<sup>94,95</sup>, Bin Zhu<sup>96</sup>, Esa Pitkänen<sup>8</sup>, Yilong Li<sup>1</sup>, Natalie Saini<sup>97</sup>, Leszek J. Klimczak<sup>98</sup>, Joachim Weischenfeld<sup>48,99,100</sup>, Nikos Sidiropoulos<sup>100</sup>, Ludmil B. Alexandrov<sup>1,101</sup>, Raquel Rabionet<sup>47,48,102</sup>, Georgia Escaramis<sup>47,103,104</sup>, Mattia Bosio<sup>40,47,49</sup>, Aliaksei T. Holik<sup>47</sup>, Hana Susak<sup>47,49</sup>, Aparna Prasad<sup>49</sup>, Serap Erkek<sup>8</sup>, Claudia Calabrese<sup>7,8</sup>, Benjamin Raeder<sup>8</sup>, Eoghan Harrington<sup>105</sup>, Simon Mayes<sup>106</sup>, Daniel Turner<sup>106</sup>, Sissel Juul<sup>105</sup>, Steven A. Roberts<sup>107</sup>, Lei Song<sup>96</sup>, Roelof Koster<sup>108</sup>, Lisa Mirabello<sup>96</sup>, Xing Hua<sup>96</sup>, Tomas J. Tanskanen<sup>109</sup>, Marta Tojo<sup>14</sup>, Jieming Chen<sup>64,110</sup>, Lauri A. Aaltonen<sup>111</sup>, Gunnar Rätsch<sup>112</sup>, Roland F. Schwarz<sup>718,119,120</sup>, Atul J. Butte<sup>121</sup>, Alvis Brazma<sup>1</sup>, Stephen J. Chanock<sup>96</sup>, Nijlankan Chatterjee<sup>122,123</sup>, Oliver Stegle<sup>78,124</sup>, Olivier Harismendy<sup>125</sup>, G. Steven Bova<sup>126</sup>, Dmitry A. Gordenin<sup>97</sup>, David Haan<sup>9</sup>, Lina Sieverling<sup>27,128</sup>, Lars Feuerbach<sup>127</sup>, Don Chalmers<sup>129</sup>, Yann Joly<sup>130</sup>, Bartha Knoppers<sup>130</sup>, Fruzsina Molnár-Gábor<sup>131</sup>, Mark Phillips<sup>130</sup>, Adrian Thorogood<sup>130</sup>, David Townsend<sup>130</sup>, Mary Goldman<sup>132</sup>, Nuno A. Fonseca<sup>713</sup>, Qian Xiang<sup>15</sup>, Brian Craft<sup>132</sup>, Elena Piñeiro-Yáñez<sup>134</sup>, Alfonso Muñoz<sup>7</sup>, Robert Petryszak<sup>7</sup>, Anja Füllgrabe<sup>7</sup>, Fatima Al-Shahrour<sup>134</sup>, María Keays<sup>7</sup>, David Haussler<sup>132,135</sup>, John Weinstein<sup>136,137</sup>, Wolfgang Huber<sup>8</sup>, Alfonso Valencia<sup>40,76</sup>, Irene Papatheodorou<sup>7</sup>, Jingchun Zhu<sup>132</sup>, Yu Fan<sup>32</sup>, David Torrents<sup>40,76</sup>, Matthias Bieg<sup>138,139</sup>, Ken Chen<sup>140</sup>, Zeechen Chong<sup>141</sup>, Kristian Cibulskis<sup>3</sup>, Roland Eils<sup>28,30,142,143</sup>, Robert S. Fulton<sup>26,27,35</sup>, Joseph L. Gelpi<sup>40,144</sup>, Santiago Gonzalez<sup>7,8</sup>, Ivo G. Gut<sup>48,74</sup>, Faraz Hach<sup>145,146</sup>, Michael Heindel<sup>28,30</sup>, Taobo Hu<sup>147</sup>, Vincent Huang<sup>12</sup>, Barbara Hutter<sup>139,148,149</sup>, Natalie Jäger<sup>28</sup>, Jongsun Jung<sup>150</sup>, Yogesh Kumar<sup>147</sup>, Christopher Lalansingh<sup>12</sup>, Ignaty Leshchiner<sup>3</sup>, Dimitri Livitz<sup>3</sup>, Eric Z. Ma<sup>147</sup>, Yosef E. Maruvka<sup>319,151</sup>, Ana Milovanovic<sup>40</sup>, Morten Muhlig Nielsen<sup>152</sup>, Nagarajan Paramasivam<sup>28,139</sup>, Jakob Skou Pedersen<sup>152,153</sup>, Montserrat Puiggrós<sup>40</sup>, S. Cenk Sahinalp<sup>146,154,155</sup>, Iman Sarrafi<sup>146,155</sup>, Chip Stewart<sup>3</sup>, Miranda D. Stobbe<sup>48,74</sup>, Jeremiah A. Wala<sup>3,6,156</sup>, Jiayin Wang<sup>27,58,157</sup>, Michael Wendt<sup>127,158,159</sup>, Johannes Werner<sup>28,160</sup>, Zhenggang Wu<sup>147</sup>, Hong Xu<sup>147</sup>, Takafumi N. Yamaguchi<sup>12</sup>, Venkata Yellapantula<sup>67,68</sup>, Brandi N. Davis-Dusenbery<sup>161</sup>, Robert L. Grossman<sup>162</sup>, Youngwook Kim<sup>163,164</sup>, Michael C. Heindel<sup>28,30</sup>, Jonathan Hinton<sup>1</sup>, David R. Jones<sup>1</sup>, Andrew Menzies<sup>1</sup>, Lucy Stebbings<sup>1</sup>, Julian M. Hess<sup>3,151</sup>, Mara Rosenberg<sup>319</sup>, Andrew J. Dunford<sup>3</sup>, Manaswi Gupta<sup>3</sup>, Marcin Imielinski<sup>165,166</sup>, Andrew Meyerov<sup>3,6,156</sup>, Rameen Beroukhim<sup>3,6,167</sup>, Jüri Reimand<sup>12,18</sup>, Priyanka Dhingra<sup>71,73</sup>, Francesco Favero<sup>168</sup>, Stefan Dentre<sup>1,65,91</sup>, Jeff Wintersinger<sup>169,170,171</sup>, Vasilisa Rudneva<sup>8</sup>, Ji Wan Park<sup>172</sup>, Eun Po Hong<sup>172</sup>, Seong Gu Heo<sup>172</sup>, André Kahles<sup>12,113,114,115,116</sup>, Kjong-Van Lehmann<sup>112,114,115,173,174</sup>, Cameron M. Soulette<sup>37</sup>, Yuichi Shiraishi<sup>39</sup>, Fenglin Liu<sup>175,176</sup>, Yao He<sup>175</sup>, Deniz Demircioğlu<sup>177,178</sup>, Natalie R. Davidson<sup>112,114,115,117,173</sup>, Liliana Greger<sup>7</sup>, Siliang Li<sup>179,180</sup>, Dongbing Liu<sup>179,180</sup>, Stefan G. Stark<sup>115,173,181,182</sup>, Fan Zhang<sup>175</sup>, Samirkumar B. Amin<sup>183,184,185</sup>, Peter Bailey<sup>186</sup>, Aurélien Chateigner<sup>15</sup>, Milana Frenkel-Morgenstern<sup>187</sup>, Yong Hou<sup>179,180</sup>, Matthew R. Huska<sup>118</sup>, Helena Kilpinen<sup>188</sup>, Fabien C. Lamaze<sup>12</sup>, Chang Li<sup>179,180</sup>, Xiaobo Li<sup>179,180</sup>, Xinyue Li<sup>179</sup>, Xingmin Liu<sup>179,180</sup>, Maximilian G. Marin<sup>37</sup>, Julia Markowski<sup>118</sup>, Tannistha Nandi<sup>189</sup>, Akinyemi I. Ojesina<sup>190,191,192</sup>, Qiang Pan-Hammarström<sup>179,193</sup>, Peter J. Park<sup>94,95</sup>, Chandra Sekhar Pedamallu<sup>3,6,167</sup>, Hong Su<sup>179,180</sup>, Patrick Tan<sup>194</sup>, Bin Tan Teh<sup>194,195,196,197,198</sup>, Jian Wang<sup>179</sup>, Heng Xiong<sup>179,180</sup>, Chen Ye<sup>179,180</sup>, Christina Yung<sup>15</sup>, Xueqiang Zhang<sup>179</sup>, Liangtao Zheng<sup>175</sup>, Shida Zhu<sup>179,180</sup>, Philip Awadalla<sup>12,13</sup>, Chad J. Creighton<sup>199</sup>, Kui Wu<sup>179,180</sup>, Huanming Yang<sup>179</sup>, Jonathan Göke<sup>177,200</sup>, Zemin Zhang<sup>175,201</sup>, Angela N. Brooks<sup>3,37,156</sup>, Matthew W Fittall<sup>91</sup>, Iñigo Martincorena<sup>1</sup>, Carlota Rubio-Perez<sup>87,88,202</sup>, Malene Juul<sup>152</sup>, Steven Schumacher<sup>3,203</sup>, Ofer Shapira<sup>3,156</sup>, David Tamborero<sup>87,89</sup>, Loris Mularoni<sup>87,89</sup>, Henrik Hornshøj<sup>152</sup>, Jordi Deu-Pons<sup>89,204</sup>, Ferran Muiños<sup>87,89</sup>, Johanna Bertl<sup>152,205</sup>, Qianyun Guo<sup>153</sup>, Abel Gonzalez-Perez<sup>87,89,206</sup>, Qian Xiang<sup>207</sup>, Wojciech Bazant<sup>7</sup>, Elisabet Barrera<sup>7</sup>, Sultan T. Al-Sedairy<sup>208</sup>, Axel Aretz<sup>209</sup>, Cindy Bell<sup>210</sup>, Miguel Betancourt<sup>211</sup>, Christiane Buchholz<sup>212</sup>, Fabien Calvo<sup>213</sup>, Christine Chomienne<sup>214</sup>, Michael Dunn<sup>215</sup>, Stuart Edmonds<sup>216</sup>, Eric Green<sup>217</sup>, Shailja Gupta<sup>218</sup>, Carolyn M. Hutter<sup>217</sup>, Karine Jegalian<sup>219</sup>, Nic Jones<sup>220</sup>, Youyong Lu<sup>221,222,223</sup>, Hitoshi Nakagama<sup>224</sup>, Gerd Nettekoven<sup>225</sup>, Laura Planko<sup>225</sup>, David Scott<sup>220</sup>, Tatsuhiro Shibata<sup>226,227</sup>, Kiyo Shimizu<sup>228</sup>,

Michael R. Stratton<sup>1</sup>, Takashi Yugawa<sup>228</sup>, Giampaolo Tortora<sup>229,230</sup>, K. VijayRaghavan<sup>218</sup>, Jean C. Zenklusen<sup>231</sup>, David Townsend<sup>232</sup>, Bartha M. Knoppers<sup>130</sup>, Brice Aminou<sup>15</sup>, Javier Bartolome<sup>40</sup>, Keith A. Borevich<sup>81,233</sup>, Rich Boyce<sup>7</sup>, Alex Buchanan<sup>38</sup>, Niall J. Byrne<sup>15</sup>, Zhaohong Chen<sup>234</sup>, Sunghoon Cho<sup>235</sup>, Wan Choi<sup>236</sup>, Peter Clapham<sup>1</sup>, Michelle T. Dow<sup>234</sup>, Lewis Jonathan Dursi<sup>12,25</sup>, Juergen Eils<sup>142,143</sup>, Claudiu Farcas<sup>234</sup>, Nodirjon Fayzullaev<sup>15</sup>, Paul Flicek<sup>7</sup>, Allison P. Heath<sup>237</sup>, Oliver Hoffmann<sup>238</sup>, Jongwhi H. Hong<sup>239</sup>, Thomas J. Hudson<sup>240,241</sup>, Daniel Hübschmann<sup>30,120,142,242,243</sup>, Sinisa Ivkovic<sup>244</sup>, Seung-Hyup Jeon<sup>236</sup>, Wei Jiao<sup>12</sup>, Rolf Kabbe<sup>28</sup>, Andre Kahles<sup>112,113,114,115,174</sup>, Jules N. A. Kerssemakers<sup>28</sup>, Hyunghwan Kim<sup>236</sup>, Jihoon Kim<sup>245</sup>, Michael Koscher<sup>246</sup>, Antonios Koures<sup>234</sup>, Milena Kovacevic<sup>244</sup>, Chris Lawerenz<sup>143</sup>, Jia Liu<sup>247</sup>, Sanja Mijalkovic<sup>244</sup>, Ana Mijalkovic Mijalkovic-Lazic<sup>244</sup>, Satoru Miyano<sup>39</sup>, Mia Nastic<sup>244</sup>, Jonathan Nicholson<sup>1</sup>, David Ocana<sup>7</sup>, Kazuhiro Ohi<sup>39</sup>, Lucila Ohno-Machado<sup>234</sup>, Todd D. Pihl<sup>248</sup>, Manuel Prinz<sup>28</sup>, Petar Radovic<sup>244</sup>, Charles Short<sup>7</sup>, Heidi J. Sofia<sup>217</sup>, Jonathan Spring<sup>162</sup>, Adam J. Struck<sup>38</sup>, Nebojsa Tijanic<sup>244</sup>, David Vicente<sup>40</sup>, Zhining Wang<sup>231</sup>, Ashley Williams<sup>234</sup>, Youngchoon Woo<sup>236</sup>, Adam J. Wright<sup>12</sup>, Liming Yang<sup>231</sup>, Mark P. Hamilton<sup>249</sup>, Todd A. Johnson<sup>233</sup>, Abdullah Kahraman<sup>250,251,252</sup>, Manolis Kellis<sup>3,253</sup>, Paz Polak<sup>3,4,6</sup>, Richard Sallari<sup>3</sup>, Nasa Sinnott-Armstrong<sup>3,4,5</sup>, Christian von Mering<sup>252,254</sup>, Sergi Beltran<sup>49,74</sup>, Daniela S. Gerhard<sup>255</sup>, Marta Gut<sup>49,74</sup>, Jean-Rémi Trotta<sup>74</sup>, Justin P. Whalley<sup>74</sup>, Beifang Niu<sup>256</sup>, Shadrille M. G. Espiritu<sup>12</sup>, Shengjie Gao<sup>179</sup>, Yi Huang<sup>157,257</sup>, Christopher M. Lalansingh<sup>12</sup>, Jon W. Teague<sup>1</sup>, Michael C. Wendt<sup>27,158,159</sup>, Federico Abascal<sup>1</sup>, Gary D. Bader<sup>13</sup>, Pratiti Bandopadhyay<sup>3,258,259</sup>, Jonathan Barenboim<sup>12</sup>, Søren Brunak<sup>260,261</sup>, Joana Carlevaro-Fita<sup>262,263,264</sup>, Dimple Chakravarty<sup>265,266</sup>, Calvin Wing Yiu Chan<sup>28,128</sup>, Jung Kyoon Choi<sup>267</sup>, Klev Diamanti<sup>268</sup>, J. Lynn Fink<sup>40,269</sup>, Joan Frigola<sup>204</sup>, Carlo Gambacorti-Passerini<sup>270</sup>, Dale W. Garsed<sup>271</sup>, Nicholas J. Haradhvala<sup>319</sup>, Arif O. Harman<sup>164,272</sup>, Mohamed Helmy<sup>170</sup>, Carl Herrmann<sup>28,30,273</sup>, Asger Hobolth<sup>153,205</sup>, Ermin Hodzic<sup>155</sup>, Chen Hong<sup>127,128</sup>, Keren Isaev<sup>12,18</sup>, Jose M. G. Izarzugaza<sup>260</sup>, Rory Johnson<sup>263,274</sup>, Randi Istrup Juul<sup>152</sup>, Jaegil Kim<sup>3</sup>, Jong K. Kim<sup>275</sup>, Jan Komorowski<sup>268,276</sup>, Andrés Lanzós<sup>263,264,274</sup>, Erik Larsson<sup>112</sup>, Donghoon Lee<sup>64</sup>, Shantao Li<sup>64</sup>, Xiaotong Li<sup>64</sup>, Ziao Lin<sup>3,277</sup>, Eric Minwei Liu<sup>71,73,278</sup>, Lucas Lovchovsky<sup>63,64,185</sup>, Shaoke Lou<sup>63,64</sup>, Tobias Madsen<sup>152</sup>, Kathleen Marchal<sup>279,280</sup>, Alexander Martinez-Fundichely<sup>71,72,73</sup>, Patrick D. McGillivray<sup>63</sup>, William Meyerson<sup>64,281</sup>, Marta Paczkowska<sup>12</sup>, Keunchil Park<sup>282,283</sup>, Kiejung Park<sup>284</sup>, Tirso Pons<sup>285</sup>, Sergio Pulido-Tamayo<sup>279,280</sup>, Iker Reyes-Salazar<sup>87</sup>, Matthew A. Reyna<sup>286</sup>, Mark A. Rubin<sup>274,287,288,289,290</sup>, Leonidas Salichos<sup>63,64</sup>, Chris Sander<sup>112,156,291,292</sup>, Steven E. Schumacher<sup>3,203</sup>, Mark Shackleton<sup>271</sup>, Ciyue Shen<sup>292,293</sup>, Raunak Shrestha<sup>146</sup>, Shimin Shua<sup>12,13</sup>, Tatsuhiko Tsunoda<sup>233,294,295,296</sup>, Husen M. Umer<sup>268,297</sup>, Liis Uusküla-Reimand<sup>288,299</sup>, Lieven P. C. Verbeke<sup>280,300</sup>, Claes Wadelius<sup>301</sup>, Lina Wadi<sup>12</sup>, Jonathan Warrell<sup>63,64</sup>, Guanming Wu<sup>302</sup>, Jun Yu<sup>303</sup>, Jing Zhang<sup>64</sup>, Xuanping Zhang<sup>157,304</sup>, Yan Zhang<sup>64,305,306</sup>, Zhongming Zhao<sup>307</sup>, Lihua Zou<sup>308</sup>, Michael S. Lawrence<sup>319,233</sup>, Benjamin J. Raphael<sup>286</sup>, Peter J. Bailey<sup>186</sup>, David Crafo<sup>3,309</sup>, Mary J. Goldman<sup>132</sup>, Hiroyuki Aburatani<sup>310</sup>, Hans Binder<sup>311,312</sup>, Huy Q. Dinh<sup>313</sup>, Simon C. Heath<sup>48,74</sup>, Steve Hoffmann<sup>311,312,314,315</sup>, Charles David Imbusch<sup>127</sup>, Helene Kretzmer<sup>312,315</sup>, Peter W. Laird<sup>316</sup>, Jose I. Martin-Subero<sup>76,317</sup>, Genta Nagae<sup>310,318</sup>, Hui Shen<sup>319</sup>, Qi Wang<sup>246</sup>, Dieter Weichenhan<sup>320</sup>, Wanding Zhou<sup>319</sup>, Benjamin P. Berman<sup>313,321,322</sup>, Benedikt Brors<sup>127,148,323</sup>, Christoph Plass<sup>320</sup>, Kadir C. Akdemir<sup>140</sup>, David D. L. Bowtell<sup>271</sup>, Kathleen H. Burns<sup>324,325</sup>, John Budanovich<sup>3,326</sup>, Kin Chan<sup>327</sup>, Ana Dueso-Barroso<sup>40</sup>, Paul A. Edwards<sup>328,329</sup>, Dariush Etemadmoghadam<sup>271</sup>, James E. Hake<sup>330</sup>, David T. W. Jones<sup>331,332</sup>, Young Seok Ju<sup>1267</sup>, Marat D. Kazanov<sup>333,334,335</sup>, Youngil Koh<sup>336,337</sup>, Kiran Kumar<sup>3</sup>, Eunjung Alice Lee<sup>338</sup>, Jake June-Koo Lee<sup>94,95</sup>, Andy G. Lynch<sup>328,329,339</sup>, Geoff Macintyre<sup>328</sup>, Florian Markowetz<sup>328,329</sup>, Fabio C. P. Navarro<sup>63</sup>, John V. Pearson<sup>240,341</sup>, Karsten Rippe<sup>120</sup>, Ralph Scully<sup>342</sup>, Izar Villasante<sup>40</sup>, Nicola Waddell<sup>1340,341</sup>, Lixing Yang<sup>343</sup>, Xiaotong Yao<sup>165,344</sup>, Sung-Soo Yoon<sup>337</sup>, Cheng-Zhong Zhang<sup>3,6,156</sup>, Erik N. Bergstrom<sup>345</sup>, Arnoud Boot<sup>193,346</sup>, Kyle Covington<sup>34</sup>, Akihiro Fujimoto<sup>61</sup>, Mi Ni Huang<sup>185,346</sup>, S. M. Ashiquil Islam<sup>101</sup>, John R. McPherson<sup>195,346</sup>, Sandro Morganello<sup>347,348,349</sup>, Alvin Wei Tian Ng<sup>350</sup>, Stephenie D. Prokopen<sup>12</sup>, Ignacio Vázquez-García<sup>167,351,352</sup>, Yang Wu<sup>195,346</sup>, Fouad Youisif<sup>12</sup>, Willie Yu<sup>353</sup>, Steven G. Rozen<sup>195,196,346</sup>, Vasilisa A. Rudneva<sup>8</sup>, Suyash S. Shringarpure<sup>45</sup>, Daniel J. Turne<sup>106</sup>, Tian Xia<sup>354</sup>, Gurnit Atwa<sup>12,13,171</sup>, David K. Chang<sup>186,355</sup>, Susanna L. Cooke<sup>186</sup>, Bishop M. Faltas<sup>117</sup>, Syed Haider<sup>12</sup>, Vera B. Kaiser<sup>356</sup>, Rosa Karic<sup>357</sup>, Mamoru Kato<sup>3</sup>, Kirsten Kübler<sup>3,6,19</sup>, Adam Margolin<sup>38</sup>, Sancha Martin<sup>1,359</sup>, Serena Nik-Zainal<sup>1360,361,362</sup>, Christine P'ng<sup>12</sup>, Colin A. Semple<sup>356</sup>, Jaclyn Smith<sup>38</sup>, Ren X. Sun<sup>12</sup>, Kevin Thai<sup>15</sup>, Derek W. Wright<sup>363,364</sup>, Ke Yuan<sup>328,359,365</sup>, Andrew V. Biankin<sup>186,355,366,367</sup>, Levi Garraway<sup>156</sup>, Sean M. Grimmond<sup>368</sup>, David J. Adams<sup>1</sup>, Pavana Anur<sup>369</sup>, Shaolong Cao<sup>32</sup>, Elizabeth L. Christie<sup>271</sup>, Marek Cmero<sup>370,371,372</sup>, Yupeng Cun<sup>373</sup>, Kevin J. Dawson<sup>1</sup>, Stefan C. Dentre<sup>1,65,91</sup>, Amit G. Deshwar<sup>374</sup>, Nilgun Donmez<sup>146,155</sup>, Ruben M. Drews<sup>328</sup>, Moritz Gerstung<sup>7,8</sup>, Gavin Ha<sup>3</sup>, Kerstin Haase<sup>91</sup>, Lara Jerman<sup>8,375</sup>, Yuan Ji<sup>376,377</sup>, Clemency Jolly<sup>91</sup>, Juhée Lee<sup>378</sup>, Henry Lee-Six<sup>1</sup>, Salem Malikic<sup>146,155</sup>, Thomas J. Mitchell<sup>1,329,379</sup>, Quidat D. Morris<sup>171,380</sup>, Layla Oesper<sup>381</sup>, Martin Seifritz<sup>37</sup>, Myron Peto<sup>382</sup>, Daniel Rosebrock<sup>3</sup>, Yulia Rubanova<sup>386,171</sup>, Adriana Salcedo<sup>12</sup>, Shunhajit Sengupta<sup>383</sup>, Ruian Shi<sup>380</sup>, Seung Jun Shin<sup>182</sup>, Oliver Spiro<sup>3</sup>, Shankar Vembu<sup>380,384</sup>, Jeffrey A. Wintersinger<sup>169,170,171</sup>, Tsun-Po Yang<sup>373</sup>, Kaixian Yu<sup>385</sup>, Hongtu Zhu<sup>386,387</sup>, Paul T. Spellman<sup>388</sup>, John N. Weinstein<sup>136,137</sup>, Yiwen Chen<sup>32</sup>, Masashi Fujita<sup>81</sup>, Leng Han<sup>304</sup>, Takanori Hasegawa<sup>39</sup>, Mitsuhiko Komura<sup>39</sup>, Jun Li<sup>32</sup>, Shinichi Mizuno<sup>389</sup>, Eigo Shimizu<sup>39</sup>, Yumeng Wang<sup>32,390</sup>, Yanxun Xu<sup>391</sup>, Rui Yamaguchi<sup>39</sup>, Fan Yang<sup>380</sup>, Yang Yang<sup>304</sup>, Christopher J. Yoon<sup>267</sup>, Yuan Yuan<sup>32</sup>, Han Liang<sup>32</sup>, Malik Alawi<sup>392,393</sup>, Ivan Borozan<sup>12</sup>, Daniel S. Brewer<sup>394,395</sup>, Colin S. Cooper<sup>395,396,397</sup>, Nikita Desai<sup>15</sup>, Adam Grundhoff<sup>292,398</sup>, Murat Iskarc<sup>399</sup>, Xiaoping Su<sup>400</sup>, Marc Zapata<sup>399</sup>, Peter Lichter<sup>148,399</sup>, Kathryn Alsop<sup>271</sup>, Timothy J. C. Bruxner<sup>269</sup>, Angelika N. Christ<sup>269</sup>, Stephen M. Cordon<sup>401</sup>, Prue A. Cowin<sup>402</sup>, Ronny Drapkin<sup>403</sup>, Sian Fereday<sup>402</sup>, Joshy George<sup>185</sup>, Anne Hamilton<sup>402</sup>, Oliver Holmes<sup>340,341</sup>, Jillian A. Hung<sup>404,405</sup>, Karin S. Kassahn<sup>269,406</sup>, Stephen H. Kazakoff<sup>402,404</sup>, Catherine J. Kennedy<sup>407,408</sup>, Conrad R. Leonard<sup>340,341</sup>, Linda Mileschkin<sup>27</sup>, David K. Miller<sup>269,355,409</sup>, Gisela Mir Arnau<sup>402</sup>, Chris Mitchell<sup>402</sup>, Felicity Newell<sup>340,341</sup>, Katia Nones<sup>340,341</sup>, Ann-Marie Patch<sup>340,341</sup>, Michael C. Quinn<sup>340,341</sup>, Darrin F. Taylor<sup>269</sup>, Heather Thorne<sup>402</sup>, Nadia Traficante<sup>402</sup>,

Ravikiran Vedururu<sup>402</sup>, Nick M. Waddell<sup>341</sup>, Paul M. Waring<sup>410</sup>, Scott Wood<sup>340,341</sup>, Qinying Xu<sup>340,341</sup>, Anna deFazio<sup>411,412,413</sup>, Matthew J. Anderson<sup>269</sup>, Davide Antonello<sup>414</sup>, Andrew P. Barbour<sup>415,416</sup>, Claudio Bassi<sup>414</sup>, Samantha Bersani<sup>417</sup>, Ivana Catedral<sup>417,418</sup>, Lorraine A. Chantrill<sup>355,419</sup>, Yoke-Eng Chiew<sup>411</sup>, Angela Chou<sup>355,420</sup>, Sara Cingarlini<sup>229</sup>, Nicole Cloonan<sup>421</sup>, Vincenzo Corbo<sup>418,422</sup>, Maria Vittoria Davi<sup>423</sup>, Fraser R. Duthie<sup>186,424</sup>, Anthony J. Gil<sup>355,420</sup>, James G. Kench<sup>355,420,427</sup>, Luca Landoni<sup>414</sup>, Rita T. Lawlor<sup>418</sup>, Andrea Maffiicini<sup>418</sup>, Neil D. Janet S. Graham<sup>186,425</sup>, Ivon Harliwong<sup>269</sup>, Nigel B. Jamieson<sup>186,367,426</sup>, Amber L. Johns<sup>355,409</sup>, Merrett<sup>414,428</sup>, Marco Miotto<sup>414</sup>, Elizabeth A. Musgrove<sup>186</sup>, Adnan M. Nagrial<sup>355</sup>, Karin A. Oien<sup>410,429</sup>, Marina Pajic<sup>355</sup>, Mark Pinese<sup>430</sup>, Alan J. Robertson<sup>269</sup>, Ilse Rooman<sup>355</sup>, Borislav C. Rusev<sup>418</sup>, Jaswinder S. Samra<sup>414,420</sup>, Maria Scardon<sup>417</sup>, Christopher J. Scarlett<sup>355</sup>, Aldo Scarpa<sup>418</sup>, Elisabetta Sereni<sup>414</sup>, Katarzyna O. Sikora<sup>418</sup>, Michele Simbolo<sup>422</sup>, Morgan L. Taschuk<sup>15</sup>, Christopher W. Toon<sup>355</sup>, Caterina Vicentini<sup>418</sup>, Jianmin Wu<sup>355</sup>, Nikolajs Zeps<sup>432,433</sup>, Andreas Behren<sup>434</sup>, Hazel Burke<sup>435</sup>, Jonathan Cebon<sup>434</sup>, Rebecca A. Dagg<sup>436</sup>, Ricardo De Paoli-Iseppi<sup>437</sup>, Ken Dutton-Regester<sup>340</sup>, Matthew A. Field<sup>438</sup>, Anna Fitzgerald<sup>439</sup>, Peter Hervey<sup>435</sup>, Valerie Jakrot<sup>435</sup>, Peter A. Johansson<sup>340</sup>, Hojabr Kakavand<sup>437</sup>, Richard F. Kefford<sup>440</sup>, Loretta M. S. Lau<sup>441</sup>, Georgina V. Long<sup>442</sup>, Hilda A. Pickett<sup>441</sup>, Antonia L. Pritchard<sup>340</sup>, Gulietta M. Pupo<sup>443</sup>, Robyn P. M. Saw<sup>442</sup>, Sarah-Jane Schramm<sup>444</sup>, Catherine A. Shang<sup>439</sup>, Ping Shang<sup>442</sup>, Andrew J. Spillane<sup>442</sup>, Jonathan R. Stretch<sup>442</sup>, Varsha Tembe<sup>411,444</sup>, John F. Thompson<sup>442</sup>, Riccardo E. Vilain<sup>445</sup>, James S. Wilmoth<sup>442</sup>, Jean Y. Cheng<sup>454</sup>, Nicholas K. Hayward<sup>340,435</sup>, Graham J. Mann<sup>411,447</sup>, Richard A. Scolyer<sup>412,442,445,448</sup>, John Bartlett<sup>449,450</sup>, Prashant Bavi<sup>451</sup>, Dianne E. Chadwick<sup>452</sup>, Michelle Chan-Seng-Yue<sup>451</sup>, Sean Cleary<sup>451,453</sup>, Ashton A. Connor<sup>453,454</sup>, Karolina Czajka<sup>241</sup>, Robert E. Denroche<sup>451</sup>, Neesha C. Dhani<sup>455</sup>, Jenna Eagles<sup>241</sup>, Steven Gallinger<sup>451,453,454</sup>, Robert C. Grant<sup>451,454</sup>, David Hedley<sup>455</sup>, Michael A. Hollingsworth<sup>456</sup>, Gun Ho Jang<sup>451</sup>, Jeremy Johns<sup>241</sup>, Sangeetha Kalimuthu<sup>451</sup>, Sheng-Ben Liang<sup>457</sup>, Ilinca Lungu<sup>451,458</sup>, Xuemeli Luo<sup>12</sup>, Faridah Mbabaali<sup>241</sup>, Treasa A. McPherson<sup>454</sup>, Jessica K. Miller<sup>241</sup>, Malcolm J. Moore<sup>455</sup>, Faiyaz Notta<sup>451,459</sup>, Danielle Pasternack<sup>241</sup>, Gloria M. Petersen<sup>460</sup>, Michael H. A. Roehrl<sup>18,451,461,462,463</sup>, Michelle Sam<sup>241</sup>, Iris Selander<sup>454</sup>, Stefano Serra<sup>410</sup>, Sagedeh Shahabi<sup>457</sup>, Sarah P. Thayer<sup>456</sup>, Lee E. Timms<sup>241</sup>, Gavin W. Wilson<sup>12,451</sup>, Julie M. Wilson<sup>451</sup>, Bradly G. Wouters<sup>464</sup>, John D. McPherson<sup>241,451,465</sup>, Timothy A. Beck<sup>15,466</sup>, Vinayak Bhandari<sup>12</sup>, Colin C. Collins<sup>14</sup>, Neil E. Fleschner<sup>467</sup>, Natalie S. Fox<sup>12</sup>, Michael Fraser<sup>12</sup>, Lawrence E. Heisler<sup>468</sup>, Emilie Lalonde<sup>12</sup>, Julie Livingstone<sup>12</sup>, Alice Meng<sup>469</sup>, Veronica Y. Sabelnykova<sup>12</sup>, Yu-Jia Shiah<sup>12</sup>, Theodorus Van der Kwast<sup>470</sup>, Robert G. Bristow<sup>18,471,472,473,474</sup>, Shuai Ding<sup>475</sup>, Daiming Fan<sup>476</sup>, Lin Li<sup>479</sup>, Yongzhan Nie<sup>476,477</sup>, Xiao Xiao<sup>157</sup>, Rui Xing<sup>222,478</sup>, Shanlin Yang<sup>475</sup>, Yingyan Yu<sup>479</sup>, Yong Zhou<sup>179</sup>, Rosamonde E. Banks<sup>480</sup>, Guillaume Bourque<sup>481,482</sup>, Paul Brennan<sup>483</sup>, Louis Letourneau<sup>484</sup>, Yasser Riazalhosseini<sup>482</sup>, Ghislaine Scelo<sup>483</sup>, Naveen Vasudev<sup>480,485</sup>, Juris Viksna<sup>486</sup>, Mark Lathrop<sup>482</sup>, Jörg Tost<sup>487</sup>, Sung-Min Ahn<sup>488</sup>, Samuel Aparicio<sup>489</sup>, Laurent Arnould<sup>490</sup>, M. R. Aure<sup>491</sup>, Shiram G. Bhosle<sup>1</sup>, Ewan Birney<sup>47</sup>, Ake Borg<sup>492</sup>, Sandrine Boyault<sup>493</sup>, Arie B. Brinkman<sup>494</sup>, Jane E. Brock<sup>495</sup>, Annegien Broeks<sup>496</sup>, Anne-Lise Børresen<sup>494</sup>, Carlos Caldas<sup>497,498</sup>, Suet-Feung Chin<sup>497,498</sup>, Helen Davies<sup>1,360,361</sup>, Christine Desmedt<sup>499,500</sup>, Luc Dirix<sup>501</sup>, Serge Dronov<sup>1</sup>, Anna Ehinger<sup>502</sup>, Jorunn E. Eyfjord<sup>503</sup>, Aquila Fatima<sup>203</sup>, John A. Foekens<sup>504</sup>, P. Andrew Futreal<sup>1505</sup>, Øystein Garred<sup>506,507</sup>, Dilip D. Giri<sup>508</sup>, Dominik Glodzik<sup>1</sup>, Dorte Grabau<sup>509</sup>, Holmfrid Hilmarsson<sup>503</sup>, Gerrit K. Hooijze<sup>510</sup>, Jocelyne Jacquemier<sup>511</sup>, Se Jin Jang<sup>512</sup>, Jon G. Jonasson<sup>503</sup>, Jos Jonkers<sup>513</sup>, Hyung-Yong Kim<sup>511</sup>, Tari A. King<sup>514,515</sup>, Stian Knappskog<sup>1,516</sup>, Gu Kong<sup>511</sup>, Savitri Krishnamurthy<sup>517</sup>, Sunil R. Lakhani<sup>518</sup>, Anita Langerød<sup>491</sup>, Denis Larsimont<sup>519</sup>, Hee Jin Lee<sup>512</sup>, Jeong-Yeon Lee<sup>520</sup>, Ming Ta Michael Lee<sup>505</sup>, Ole Christian Lingjaerde<sup>521</sup>, Gaetan MacGrogan<sup>522</sup>, John W. M. Martens<sup>504</sup>, Sarah O'Meara<sup>1</sup>, Iris Paupeurt<sup>524</sup>, Sarah Pinder<sup>523</sup>, Xavier Pivot<sup>524</sup>, Elena Provenzano<sup>525</sup>, Colin A. Purdie<sup>526</sup>, Manasa Ramakrishna<sup>1</sup>, Kamna Ramakrishnan<sup>1</sup>, Jorge Reis-Filho<sup>508</sup>, Andrea L. Richardson<sup>203</sup>, Markus Ringné<sup>492</sup>, Javier Bartolomé Rodríguez<sup>40</sup>, F. Germán Rodríguez-González<sup>261</sup>, Gilles Romieu<sup>527</sup>, Roberto Salgado<sup>40</sup>, Torill Sauer<sup>521</sup>, Rebecca Shepherd<sup>1</sup>, Anieta M. Sieuwerts<sup>504</sup>, Peter T. Simpson<sup>518</sup>, Marcel Smid<sup>504</sup>, Christos Sotiriou<sup>234</sup>, Paul N. Span<sup>528</sup>, Ólafur Andri Stefánsson<sup>529</sup>, Alasdair Stenhouse<sup>530</sup>, Henk G. Stunnenberg<sup>180,531</sup>, Fred Sweep<sup>532</sup>, Benita Kiat Tee Tan<sup>533</sup>, Gilles Thomas<sup>534</sup>, Alastair M. Thompson<sup>530</sup>, Stefania Tommasi<sup>535</sup>, Isabelle Treilleux<sup>536,537</sup>, Andrew Tutt<sup>203</sup>, Naoto T. Ueno<sup>387</sup>, Steven Van Laere<sup>501</sup>, Gert G. Van den Eynden<sup>501</sup>, Peter Vermeulen<sup>501</sup>, Alain Viari<sup>418</sup>, Anne Vincent-Salomon<sup>531</sup>, Bernice H. Wong<sup>538</sup>, Lucy Yates<sup>1</sup>, Xueqing Zou<sup>1</sup>, Carolien H. M. van Deuren<sup>539</sup>, Marc J. van de Vijver<sup>410</sup>, Laura van't Veer<sup>540</sup>, Ole Ammerpoh<sup>541,542,543</sup>, Sietse Aukema<sup>542,543,544</sup>, Anke K. Bergmann<sup>545</sup>, Stephan H. Bernhart<sup>311,312,315</sup>, Arndt Borkhardt<sup>546</sup>, Christoph Bors<sup>547</sup>, Birgit Burkhardt<sup>548</sup>, Alexander Claviez<sup>549</sup>, Maria Elisabeth Goebler<sup>550</sup>, Andrea Haake<sup>541</sup>, Siegfried Haas<sup>547</sup>, Martin Hansmann<sup>551</sup>, Jessica I. Hoell<sup>546</sup>, Michael Hummel<sup>552</sup>, Dennis Karsch<sup>553</sup>, Wolfram Klapper<sup>544</sup>, Michael Kneba<sup>553</sup>, Markus Kreuz<sup>554</sup>, Dieter Kube<sup>555</sup>, Ralf Küppers<sup>556</sup>, Dido Lenze<sup>552</sup>, Markus Loeffler<sup>554</sup>, Cristina López<sup>80,551</sup>, Luisa Mantovani-Löffler<sup>557</sup>, Peter Möller<sup>558</sup>, German Ott<sup>559</sup>, Bernhard Radlwimmer<sup>399</sup>, Julia Richter<sup>541,544</sup>, Marius Rohde<sup>560</sup>, Philip C. Rosenstiel<sup>561</sup>, Andreas Rosenwald<sup>562</sup>, Markus B. Schilhabel<sup>561</sup>, Stefan Schreiber<sup>563</sup>, Peter F. Stadler<sup>311,312,315</sup>, Peter Staib<sup>564</sup>, Stephan Stilgenbauer<sup>565</sup>, Stephanie Sungalee<sup>8</sup>, Monika Szczepanowski<sup>544</sup>, Umut H. Toprak<sup>30,566</sup>, Lorenz H. P. Trümper<sup>555</sup>, Rabea Wagener<sup>80,541</sup>, Thorsten Zenz<sup>149</sup>, Volker Hovestadt<sup>399</sup>, Christof von Kalle<sup>120</sup>, Marco Kool<sup>246,331</sup>, Andrey Korshunov<sup>246</sup>, Pablo Landgraf<sup>567,568</sup>, Hans Lehrach<sup>569</sup>, Paul A. Northcott<sup>570</sup>, Stefan M. Pfister<sup>246,331,571</sup>, Guido Reifenberger<sup>568</sup>, Hans-Jörg Warnatz<sup>569</sup>, Stephan Wolf<sup>572</sup>, Marie-Laure Yaspo<sup>569</sup>, Yassen Assenov<sup>573</sup>, Clarissa Gerhauser<sup>320</sup>, Sarah Minner<sup>574</sup>, Thorsten Schlömm<sup>99,575</sup>, Ronald Simon<sup>576</sup>, Guido Sauter<sup>576</sup>, Holger Sültmann<sup>149,577</sup>, Nidhan K. Biswas<sup>578</sup>, Arindam Maitra<sup>578</sup>, Partha P. Majumder<sup>578</sup>, Rajiv Sarin<sup>579</sup>, Stefano Barbi<sup>422</sup>, Giada Bonizzato<sup>418</sup>, Cinzia Cantù<sup>418</sup>, Angelo P. Dei Tos<sup>580</sup>, Matteo Fassan<sup>581</sup>, Sonia Grimaldi<sup>418</sup>, Claudio Luchini<sup>417</sup>, Giuseppe Mallo<sup>414</sup>, Giovanni Marchegiani<sup>414</sup>, Michele Milella<sup>29</sup>, Salvatore Paiella<sup>414</sup>, Antonio Pea<sup>414</sup>, Paolo Pederzoli<sup>414</sup>, Andrea Ruzzenente<sup>414</sup>, Roberto Salvia<sup>414</sup>, Nicola Sperandio<sup>418</sup>, Yasuhito Arai<sup>226</sup>, Natsuko Hama<sup>226</sup>, Nobuyoshi Hiraoka<sup>582</sup>

Fumie Hosoda<sup>226</sup>, Hiromi Nakamura<sup>226</sup>, Hidenori Ojima<sup>583</sup>, Takuji Okusaka<sup>584</sup>, Yasushi Totoki<sup>226</sup>, Tomoko Urushidate<sup>227</sup>, Masashi Fukayama<sup>585</sup>, Shumpei Ishikawa<sup>586</sup>, Hitoshi Katai<sup>587</sup>, Hiroto Katoh<sup>586</sup>, Daisuke Komura<sup>586</sup>, Hiroyumi Rokutan<sup>588</sup>, Mihoko Saito-Adachi<sup>588</sup>, Akihiro Suzuki<sup>310,588</sup>, Hirokazu Taniguchi<sup>589</sup>, Kenji Tatsuno<sup>310</sup>, Tetsuo Ushiku<sup>585</sup>, Shinichi Yachida<sup>226,590</sup>, Shogo Yamamoto<sup>310</sup>, Hiroshi Aikata<sup>591</sup>, Koji Arihiro<sup>591</sup>, Shun-ichi Arizumi<sup>592</sup>, Kazuaki Chayama<sup>591</sup>, Mayuko Furuta<sup>81</sup>, Kunihito Gotoh<sup>593</sup>, Shinya Hayami<sup>594</sup>, Satoshi Hirano<sup>595</sup>, Yoshihiko Kawakami<sup>591</sup>, Kazuhiro Maejima<sup>81</sup>, Toru Nakamura<sup>595</sup>, Kaeoru Nakano<sup>81</sup>, Hideki Ohdan<sup>591</sup>, Aya Sasaki-Oku<sup>81</sup>, Hiroko Tanaka<sup>39</sup>, Masaki Ueno<sup>594</sup>, Masakazu Yamamoto<sup>592</sup>, Hiroki Yamaue<sup>594</sup>, Su Pin Choo<sup>596</sup>, Ioana Cutcutache<sup>195,346</sup>, Narong Khuntikeo<sup>414,597</sup>, Choon Kiat Ong<sup>598</sup>, Chawalit Pairojkul<sup>410</sup>, Irinel Popescu<sup>599</sup>, Keun Soo Ahn<sup>600</sup>, Marta Aymerich<sup>601</sup>, Armando Lopez-Guillermo<sup>602</sup>, Carlos López-Otin<sup>603</sup>, Xose S. Puente<sup>603</sup>, Elias Campo<sup>604,605</sup>, Fernanda Amary<sup>606</sup>, Daniel Baumhoer<sup>607</sup>, Sam Behjati<sup>1</sup>, Bodil Bjerkehagen<sup>607,608</sup>, P. A. Futreal<sup>505</sup>, Ola Myklebos<sup>616</sup>, Nischalan Pillay<sup>609</sup>, Patrick Tarpey<sup>610</sup>, Roberto Tirabosco<sup>611</sup>, Olga Zaikova<sup>612</sup>, Adrienne M. Flanagan<sup>613</sup>, Jacqueline Boulwtow<sup>614</sup>, David T. Bowen<sup>1</sup>, Mario Cazzola<sup>615</sup>, Anthony R. Green<sup>239</sup>, Eva Hellstrom-Lindberg<sup>616</sup>, Luca Malcovati<sup>615</sup>, Jyoti Nangalia<sup>617</sup>, Elli Papaemmanuil<sup>1</sup>, Paresh Vyas<sup>340,618</sup>, Yeng Ang<sup>619</sup>, Hugh Bar<sup>620</sup>, Duncan Beardsmore<sup>621</sup>, Matthew Eldridge<sup>628</sup>, James Gossage<sup>622</sup>, Nicola Grehan<sup>361</sup>, George B. Hanna<sup>623</sup>, Stephen J. Hayes<sup>624,625</sup>, Ted R. Hupp<sup>627</sup>, David Khoo<sup>627</sup>, Jesper Lagergren<sup>616,628</sup>, Laurence B. Lovat<sup>188</sup>, Shona MacRae<sup>136</sup>, Maria O'Donovan<sup>361</sup>, J. Robert O'Neill<sup>629</sup>, Simon L. Parsons<sup>630</sup>, Shaun R. Preston<sup>631</sup>, Sonia Puig<sup>632</sup>, Tom Roques<sup>633</sup>, Grant Sanders<sup>24</sup>, Sharmila Sothi<sup>634</sup>, Simon Tavaré<sup>638</sup>, Olga Tucker<sup>635</sup>, Richard Turkington<sup>636</sup>, Timothy J. Underwood<sup>637</sup>, Ian Welch<sup>638</sup>, Rebecca C. Fitzgerald<sup>361</sup>, Daniel M. Berney<sup>639</sup>, Johann S. De Bono<sup>396</sup>, Declan Cahill<sup>640</sup>, Niedzica Camacho<sup>396</sup>, Nening M. Dennis<sup>640</sup>, Tim Dudderidge<sup>640,641</sup>, Sandra E. Edwards<sup>396</sup>, Cyril Fisher<sup>640</sup>, Christopher S. Foster<sup>642,643</sup>, Mohammed Ghori<sup>1</sup>, Pelvender Gill<sup>618</sup>, Vincent J. Gnanapragasam<sup>279,644</sup>, Gunes Gundem<sup>278</sup>, Freddie C. Hamdy<sup>645</sup>, Steve Hawkins<sup>328</sup>, Steven Hazel<sup>1640</sup>, William Howat<sup>379</sup>, William B. Isaacs<sup>646</sup>, Katalin Kaszsi<sup>618</sup>, Jonathan D. Kay<sup>188</sup>, Vincent Knoch<sup>640</sup>, Zsolt Kote-Jarai<sup>396</sup>, Barbara Kremer<sup>1</sup>, Pardeep Kumar<sup>640</sup>, Adam Lambert<sup>618</sup>, Daniel A. Leongamornlert<sup>1,396</sup>, Naomi Livni<sup>640</sup>, Yong-Jie Lu<sup>639,647</sup>, Hayley J. Luxton<sup>188</sup>, Luke Marsden<sup>618</sup>, Charlie E. Massie<sup>328</sup>, Lucy Matthews<sup>396</sup>, Erik Mayer<sup>640,648</sup>, Ultan McDermott<sup>1</sup>, Sue Merson<sup>396</sup>, David E. Neal<sup>328,379</sup>, Anthony Ng<sup>649</sup>, David Nicol<sup>640</sup>, Christopher Ogden<sup>640</sup>, Edward W. Rowe<sup>640</sup>, Nimish C. Shah<sup>379</sup>, Sarah Thomas<sup>640</sup>, Alan Thompson<sup>640</sup>, Clare Verrill<sup>618,650</sup>, Tapio Visakorpi<sup>126</sup>, Anne Y. Warren<sup>379,651</sup>, Hayley C. Whitaker<sup>188</sup>, Hongwei Zhang<sup>647</sup>, Nicholas van As<sup>640</sup>, Rosalind A. Eeles<sup>396,640</sup>, Adam Abeshouse<sup>278</sup>, Nishant Agrawal<sup>12</sup>, Rehan Akbani<sup>361,652</sup>, Hikmat Al-Ahmadie<sup>78</sup>, Monique Albert<sup>450</sup>, Kenneth Alldape<sup>400,653</sup>, Adrian Ali<sup>654</sup>, Elizabeth L. Appelbaum<sup>271,88</sup>, Joshua Armenia<sup>655</sup>, Sylvia Asa<sup>630,656</sup>, J. Todd Auman<sup>657</sup>, Miruna Balasundaram<sup>654</sup>, Saianand Balu<sup>24</sup>, Jill Barnholtz-Sloan<sup>658,659</sup>, Oliver F. Bathe<sup>660,661</sup>, Stephen B. Baylin<sup>123,641</sup>, Christopher Benz<sup>662</sup>, Andrew Berchuck<sup>663</sup>, Mario Berrios<sup>660</sup>, Darell Bigner<sup>665</sup>, Michael Birrer<sup>19</sup>, Tom Bodenheimer<sup>24</sup>, Lori Boice<sup>632</sup>, Moiz S. Bootwalla<sup>664</sup>, Marcus Bosenberg<sup>666</sup>, Reanne Bowlby<sup>654</sup>, Jeffrey Boyd<sup>667</sup>, Russell R. Broadus<sup>400</sup>, Malcolm Brock<sup>668</sup>, Denise Brooks<sup>654</sup>, Susan Bullman<sup>3167</sup>, Samantha J. Caesar-Johnson<sup>231</sup>, Thomas E. Carey<sup>669</sup>, Rebecca Carlsen<sup>654</sup>, Robert Cerfolio<sup>670</sup>, Vishal S. Chandan<sup>671</sup>, Hsiao-Wei Chen<sup>619,655</sup>, Andrew D. Cherniack<sup>3156,167</sup>, Jeremy Chien<sup>672</sup>, Juck Cho<sup>3</sup>, Eric Chuah<sup>654</sup>, Carrie Cibulskis<sup>3</sup>, Leslie Cope<sup>673</sup>, Matthew G. Cordes<sup>27,633</sup>, Erin Curley<sup>674</sup>, Bogdan Czerniak<sup>400,627</sup>, Ludmila Danilova<sup>673</sup>, Ian J. Davis<sup>675</sup>, Timothy Defreitas<sup>3</sup>, John A. Demchok<sup>231</sup>, Noreen Dhallia<sup>654</sup>, Rajiv Dhir<sup>676</sup>, HarshaVardhan Doddapaneni<sup>34</sup>, Adel El-Naggar<sup>400,627</sup>, Ina Felau<sup>231</sup>, Martin L. Ferguson<sup>677</sup>, Gaetano Finocchiaro<sup>678</sup>, Kwun M. Fong<sup>679</sup>, Scott Frazer<sup>3</sup>, William Friedman<sup>680</sup>, Catrina C. Fronick<sup>27,633</sup>, Lucinda A. Fulton<sup>27</sup>, Stacey B. Gabriel<sup>3</sup>, Jianjiang Gao<sup>655</sup>, Nils Gehlenborg<sup>3,681</sup>, Jeffrey E. Gershenwald<sup>682,683</sup>, Ronald Ghossein<sup>708</sup>, Nasra H. Giama<sup>684</sup>, Richard A. Gibbs<sup>34</sup>, Carmen Gomez<sup>685</sup>, Ramaswamy Govindan<sup>26</sup>, D. Neil Hayes<sup>24,686,687</sup>, Apurva M. Hegde<sup>136,137</sup>, David I. Heiman<sup>3</sup>, Zachary Heins<sup>278</sup>, Austin J. Hepperla<sup>24</sup>, Andrea Holbrook<sup>664</sup>, Robert A. Holt<sup>654</sup>, Alan P. Hoyle<sup>24</sup>, Ralph H. Hruban<sup>673</sup>, Jianhong Hu<sup>34</sup>, Mei Huang<sup>632</sup>, David Huntsman<sup>688</sup>, Jason Huse<sup>278</sup>, Christine A. Iacobuzio-Donahue<sup>508</sup>, Michael Ittmann<sup>689,690</sup>, Joy C. Jayaseelan<sup>34</sup>, Stuart R. Jefferys<sup>24</sup>, Corbin D. Jones<sup>691</sup>, Steven J. M. Jones<sup>692</sup>, Hartmut Juhl<sup>693</sup>, Koo Jeong Kang<sup>694</sup>, Beth Karlan<sup>695</sup>, Katayoon Kasaian<sup>692</sup>, Electron Kebebew<sup>696,697</sup>, Hark Kyun Kim<sup>698</sup>, Viktoriya Korchina<sup>34</sup>, Ritika Kundra<sup>619,655</sup>, Phillip H. La<sup>664</sup>, Eric Lander<sup>3</sup>, Xuan Le<sup>699</sup>, Darlene Lee<sup>654</sup>, Douglas A. Levine<sup>278,700</sup>, Lora Lewis<sup>34</sup>, Tim Ley<sup>701</sup>, Haiyan Irene Li<sup>654</sup>, Pei Lin<sup>3</sup>, W. M. Linehan<sup>702</sup>, Fei Fei Liu<sup>280</sup>, Yiling Liu<sup>137</sup>, Lisa Lype<sup>703</sup>, Yussanne Ma<sup>654</sup>, Dennis T. Maglinte<sup>664,704</sup>, Elaine R. Mardis<sup>27,667,705</sup>, Jeffrey Marks<sup>414,706</sup>, Marco A. Marra<sup>654</sup>, Thomas J. Matthew<sup>37</sup>, Michael Mayo<sup>654</sup>, Karen McCune<sup>707</sup>, Samuel R. Meier<sup>3</sup>, Shaowu Meng<sup>24</sup>, Piotr A. Mieczkowski<sup>123</sup>, Tom Mikkelson<sup>708</sup>, Christopher A. Miller<sup>27</sup>, Gordon B. Mills<sup>709</sup>, Richard A. Moore<sup>654</sup>, Carl Morrison<sup>410,710</sup>, Lisle E. Mose<sup>24</sup>, Catherine D. Moser<sup>684</sup>, Andrew J. Mungall<sup>654</sup>, Karen Mungall<sup>654</sup>, David Mutch<sup>711</sup>, Donna M. Muzny<sup>712</sup>, Jerome Myers<sup>713</sup>, Yulia Newton<sup>37</sup>, Michael S. Noble<sup>3</sup>, Peter O'Donnell<sup>714</sup>, Brian Patrick O'Neill<sup>715</sup>, Angelica Ochoa<sup>278</sup>, Joong-Won Park<sup>716</sup>, Joel S. Parker<sup>717</sup>, Harvey Pass<sup>718</sup>, Alessandro Pastore<sup>112</sup>, Nathan A. Pennell<sup>1719</sup>, Charles M. Perou<sup>720</sup>, Nicholas Petrelli<sup>721</sup>, Olga Potapova<sup>722</sup>, Janet S. Rader<sup>723</sup>, Suresh Ramalingam<sup>724</sup>, W. Kimryn Rathmell<sup>1725</sup>, Victor Reuter<sup>508</sup>, Sheila M. Reynolds<sup>703</sup>, Matthew Ringel<sup>726</sup>, Jeffrey Roach<sup>727</sup>, Lewis R. Roberts<sup>684</sup>, A. Gordon Robertson<sup>654</sup>, Sara Sadeghi<sup>654</sup>, Charles Saller<sup>728</sup>, Francisco Sanchez-Vega<sup>618,655</sup>, Dirk Schadendorf<sup>148,655</sup>, Jacqueline E. Schein<sup>654</sup>, Heather K. Schmidt<sup>27</sup>, Nikolaus Schultz<sup>655</sup>, Raja Seethala<sup>720</sup>, Yasin Sebnabaoglu<sup>112</sup>, Troy Shelton<sup>674</sup>, Yan Shi<sup>24</sup>, Juliann Shih<sup>3,167</sup>, Ilya Shmulevich<sup>703</sup>, Craig Shriver<sup>731</sup>, Sabina Signoretti<sup>167,263,732</sup>, Janae V. Simons<sup>24</sup>, Samuel Singer<sup>414,733</sup>, Payal Sipahimalani<sup>654</sup>, Tara J. Skelly<sup>23</sup>, Karen Smith-McCune<sup>707</sup>, Nicholas D. Socci<sup>112</sup>, Matthew G. Soloway<sup>717</sup>, Anil K. Sood<sup>734</sup>, Angela Tam<sup>654</sup>, Donghui Tan<sup>23</sup>, Roy Tarnuzzer<sup>231</sup>, Nina Thiessen<sup>654</sup>, R. Houston Thompson<sup>735</sup>, Leigh B. Thorne<sup>632</sup>, Ming Tsao<sup>630,656</sup>, Christopher Umbricht<sup>224,621,736</sup>, David J. Van Den Berg<sup>664</sup>, Erwin G. Van Meir<sup>737</sup>, Umadevi Veluvolu<sup>23</sup>, Douglas Voet<sup>3</sup>, Linghua Wang<sup>34</sup>, Paul Weinberger<sup>738</sup>

# Article

**Daniel J. Weisenberger<sup>664</sup>, Dennis Wigle<sup>739</sup>, Matthew D. Wilkerson<sup>23</sup>, Richard K. Wilson<sup>27,740</sup>, Boris Winterhoff<sup>741</sup>, Maciej Wiznerowicz<sup>742,743</sup>, Tina Wong<sup>27,654</sup>, Wingham Wong<sup>744</sup>, Liu Xi<sup>34</sup>, Christina Yau<sup>662</sup>, Hailei Zhang<sup>3</sup>, Hongxin Zhang<sup>655</sup> & Jiashan Zhang<sup>231</sup>**

<sup>1</sup>Wellcome Sanger Institute, Hinxton, UK. <sup>2</sup>Department of Haematology, University of Cambridge, Cambridge, UK. <sup>3</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>4</sup>Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. <sup>5</sup>Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. <sup>6</sup>Harvard Medical School, Boston, MA, USA. <sup>7</sup>European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. <sup>8</sup>European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. <sup>9</sup>Biomolecular Engineering Department, University of California Santa Cruz, Santa Cruz, CA, USA. <sup>10</sup>Adaptive Oncology Initiative, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>11</sup>International Cancer Genome Consortium (ICGC)/ICGC Accelerating Research in Genomic Oncology (ICGC-ARGO) Secretariat, Toronto, Ontario, Canada. <sup>12</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>13</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. <sup>14</sup>Department of Radiation Oncology, University of California San Francisco, San Francisco, CA, USA. <sup>15</sup>Genome Informatics Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>16</sup>Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada. <sup>17</sup>Genome Informatics, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>18</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. <sup>19</sup>Massachusetts General Hospital, Boston, MA, USA. <sup>20</sup>Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. <sup>21</sup>University of California Los Angeles, Los Angeles, CA, USA. <sup>22</sup>Department of Pathology, Department of Genomic Medicine and Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>23</sup>Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>24</sup>Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>25</sup>The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>26</sup>Alvin J. Siteman Cancer Center, Washington University School of Medicine, St Louis, MO, USA. <sup>27</sup>The McDonnell Genome Institute, Washington University, St Louis, MO, USA. <sup>28</sup>Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>29</sup>Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center, Heidelberg, Germany. <sup>30</sup>Institute of Pharmacy and Molecular Biotechnology, and BioQuant, Heidelberg University, Heidelberg, Germany. <sup>31</sup>Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>32</sup>Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>33</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. <sup>34</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. <sup>35</sup>Department of Genetics and Department of Medicine, Washington University in St Louis, St Louis, MO, USA. <sup>36</sup>Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. <sup>37</sup>University of California Santa Cruz, Santa Cruz, CA, USA. <sup>38</sup>Computational Biology Program, Oregon Health & Science University, Portland, OR, USA. <sup>39</sup>The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>40</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain. <sup>41</sup>Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway. <sup>42</sup>Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>43</sup>Department of Zoology, Genetics and Physical Anthropology, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>44</sup>The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. <sup>45</sup>Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. <sup>46</sup>Annai Systems, Carlsbad, CA, USA. <sup>47</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. <sup>48</sup>Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. <sup>49</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>50</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. <sup>51</sup>Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. <sup>52</sup>Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland. <sup>53</sup>Department of Ophthalmology, Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA. <sup>54</sup>Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>55</sup>Department of Veterinary Medicine, Transmissible Cancer Group, University of Cambridge, Cambridge, UK. <sup>56</sup>Department of Biochemistry, College of Medicine, Ewha Womans University, Seoul, South Korea. <sup>57</sup>Division of Oncology, Washington University School of Medicine, St Louis, MO, USA. <sup>58</sup>School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. <sup>59</sup>The First Affiliated Hospital, Xi'an Jiaotong University, Xi'an, China. <sup>60</sup>Independent Consultant, Wellesley, MA, USA. <sup>61</sup>Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>62</sup>Biobyte Solutions, Heidelberg, Germany. <sup>63</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. <sup>64</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. <sup>65</sup>Big Data Institute, Li Ka Shing Centre, University of Oxford, Oxford, UK. <sup>66</sup>Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK. <sup>67</sup>Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>68</sup>The McDonnell Genome Institute at Washington University School of Medicine, and Department of Genetics and Department of Medicine, Siteman Cancer Center, Washington University in St Louis, St Louis, MO, USA.

<sup>69</sup>Department of Computer Science, Yale University, New Haven, CT, USA. <sup>70</sup>Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. <sup>71</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. <sup>72</sup>Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. <sup>73</sup>Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. <sup>74</sup>CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. <sup>75</sup>Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>76</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. <sup>77</sup>Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. <sup>78</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. <sup>79</sup>Human Genetics, University of Kiel, Kiel, Germany. <sup>80</sup>Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany. <sup>81</sup>RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>82</sup>Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK. <sup>83</sup>Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK. <sup>84</sup>Quantitative Genomics Laboratories (qGenomics), Barcelona, Spain. <sup>85</sup>Sage Bionetworks, Seattle, WA, USA. <sup>86</sup>Department of Biochemistry and Molecular Medicine, University of Montreal, Montreal, Quebec, Canada. <sup>87</sup>Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. <sup>88</sup>National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India. <sup>89</sup>Research Program on Biomedical Informatics, Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>90</sup>Broad Institute of Harvard and MIT, Cambridge, MA, USA. <sup>91</sup>The Francis Crick Institute, London, UK. <sup>92</sup>University of Leuven, Leuven, Belgium. <sup>93</sup>Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. <sup>94</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>95</sup>Ludwig Center at Harvard Medical School, Boston, MA, USA. <sup>96</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>97</sup>Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. <sup>98</sup>Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. <sup>99</sup>Department of Urology, Charité Universitätsmedizin Berlin, Berlin, Germany. <sup>100</sup>Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. <sup>101</sup>Department of Bioengineering and Department of Cellular and Molecular Medicine, Moores Cancer Center, University of California San Diego, La Jolla, CA, USA. <sup>102</sup>Department of Genetics, Microbiology and Statistics, University of Barcelona, IRSJD, IBUB, Barcelona, Spain. <sup>103</sup>CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. <sup>104</sup>Research Group on Statistics, Econometrics and Health (GRECS), UdG, Barcelona, Spain. <sup>105</sup>Oxford Nanopore Technologies, New York, NY, USA. <sup>106</sup>Applications Department, Oxford Nanopore Technologies, Oxford, UK. <sup>107</sup>School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA, USA. <sup>108</sup>Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>109</sup>Department of Medical and Clinical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland. <sup>110</sup>Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT, USA. <sup>111</sup>Applied Tumor Genomics Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland. <sup>112</sup>Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>113</sup>Department of Biology, ETH Zurich, Zurich, Switzerland. <sup>114</sup>Department of Computer Science, ETH Zurich, Zurich, Switzerland. <sup>115</sup>SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. <sup>116</sup>University Hospital Zurich, Zurich, Switzerland. <sup>117</sup>Weill Cornell Medical College, New York, NY, USA. <sup>118</sup>Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. <sup>119</sup>German Cancer Consortium (DKTK), Partner site Berlin, Berlin, Germany. <sup>120</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>121</sup>Baker Computational Health Sciences Institute and Department of Pediatrics, University of California, San Francisco, CA, USA. <sup>122</sup>Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. <sup>123</sup>Department of Oncology, The Johns Hopkins School of Medicine, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, MD, USA. <sup>124</sup>Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>125</sup>Department of Medicine and Moores Cancer Center, Division of Biomedical Informatics, UC San Diego School of Medicine, San Diego, CA, USA. <sup>126</sup>Faculty of Medicine and Health Technology, Tampere University and Tays Cancer Center, Tampere University Hospital, Tampere, Finland. <sup>127</sup>Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>128</sup>Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. <sup>129</sup>Centre for Law and Genetics, University of Tasmania, Hobart, Tasmania, Australia. <sup>130</sup>Centre of Genomics and Policy, McGill University and Génomique Québec Innovation Centre, Montreal, Quebec, Canada. <sup>131</sup>Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany. <sup>132</sup>UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. <sup>133</sup>CIBIO/InBIO, Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão, Portugal. <sup>134</sup>Bioinformatics Unit, Spanish National Cancer Research Center (CNIO), Madrid, Spain. <sup>135</sup>Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA. <sup>136</sup>Cancer Unit, MRC University of Cambridge, Cambridge, UK. <sup>137</sup>Department of Bioinformatics and Computational Biology and Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>138</sup>Center for Digital Health, Berlin Institute of Health (BIH) and Charité-Universitätsmedizin Berlin, Berlin, Germany. <sup>139</sup>Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer



Research Center (DKFZ), Heidelberg, Germany. <sup>140</sup>Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>141</sup>Department of Genetics and Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, USA. <sup>142</sup>Heidelberg University, Heidelberg, Germany. <sup>143</sup>New BIH Digital Health Center, Berlin Institute of Health (BIH) and Charité-Universitätsmedizin Berlin, Berlin, Germany. <sup>144</sup>Department of Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain. <sup>145</sup>Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada. <sup>146</sup>Vancouver Prostate Centre, Vancouver, British Columbia, Canada. <sup>147</sup>Division of Life Science and Applied Genomics Center, Hong Kong University of Science and Technology, Hong Kong, China. <sup>148</sup>German Cancer Consortium (DKTK), Heidelberg, Germany. <sup>149</sup>National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany. <sup>150</sup>Genome Integration Data Center, Syntekabio, Daejeon, South Korea. <sup>151</sup>Massachusetts General Hospital Center for Cancer Research, Charlestown, MA, USA. <sup>152</sup>Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark. <sup>153</sup>Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark. <sup>154</sup>Indiana University, Bloomington, IN, USA. <sup>155</sup>Simon Fraser University, Burnaby, British Columbia, Canada. <sup>156</sup>Dana-Farber Cancer Institute, Boston, MA, USA. <sup>157</sup>School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. <sup>158</sup>Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. <sup>159</sup>Department of Mathematics, Washington University in St Louis, St Louis, MO, USA. <sup>160</sup>Department of Biological Oceanography, Leibniz Institute of Baltic Sea Research, Rostock, Germany. <sup>161</sup>Seven Bridges Genomics, Charlestown, MA, USA. <sup>162</sup>University of Chicago, Chicago, IL, USA. <sup>163</sup>Department of Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul, South Korea. <sup>164</sup>Samsung Genome Institute, Seoul, South Korea. <sup>165</sup>New York Genome Center, New York, NY, USA. <sup>166</sup>Weill Cornell Medicine, New York, NY, USA. <sup>167</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>168</sup>Rigshospitalet, Copenhagen, Denmark. <sup>169</sup>Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. <sup>170</sup>The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. <sup>171</sup>Vector Institute, Toronto, Ontario, Canada. <sup>172</sup>Department of Medical Genetics, College of Medicine, Hallym University, Chuncheon, South Korea. <sup>173</sup>Department of Biology, ETH Zurich, Zurich, Switzerland. <sup>174</sup>University Hospital Zurich, Zurich, Switzerland. <sup>175</sup>Peking University, Beijing, China. <sup>176</sup>School of Life Sciences, Peking University, Beijing, China. <sup>177</sup>Computational and Systems Biology, Genome Institute of Singapore, Singapore, Singapore. <sup>178</sup>School of Computing, National University of Singapore, Singapore, Singapore. <sup>179</sup>BGI-Shenzhen, Shenzhen, China. <sup>180</sup>China National GeneBank-Shenzhen, Shenzhen, China. <sup>181</sup>Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>182</sup>Korea University, Seoul, South Korea. <sup>183</sup>Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>184</sup>Quantitative & Computational Biosciences Graduate Program, Baylor College of Medicine, Houston, TX, USA. <sup>185</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. <sup>186</sup>Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Bearsden, UK. <sup>187</sup>The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. <sup>188</sup>University College London, London, UK. <sup>189</sup>Genome Institute of Singapore, Singapore, Singapore. <sup>190</sup>Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. <sup>191</sup>HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. <sup>192</sup>O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA. <sup>193</sup>Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. <sup>194</sup>Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. <sup>195</sup>Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore. <sup>196</sup>SingHealth, Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore, Singapore. <sup>197</sup>Institute of Molecular and Cell Biology, Singapore, Singapore. <sup>198</sup>Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, Singapore, Singapore. <sup>199</sup>Department of Medicine, Baylor College of Medicine, Houston, TX, USA. <sup>200</sup>National Cancer Centre Singapore, Singapore, Singapore. <sup>201</sup>BIOPIC, ICG and College of Life Sciences, Peking University, Beijing, China. <sup>202</sup>Val d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. <sup>203</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>204</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. <sup>205</sup>Department of Mathematics, Aarhus University, Aarhus, Denmark. <sup>206</sup>Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, Spain. <sup>207</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>208</sup>King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia. <sup>209</sup>DLR Project Management Agency, Bonn, Germany. <sup>210</sup>Genome Canada, Ottawa, Ontario, Canada. <sup>211</sup>Instituto Carlos Slim de la Salud, Mexico City, Mexico. <sup>212</sup>Federal Ministry of Education and Research, Berlin, Germany. <sup>213</sup>Institut Gustave Roussy, Villejuif, France. <sup>214</sup>Institut National du Cancer (INCA), Boulogne-Billancourt, France. <sup>215</sup>The Wellcome Trust, London, UK. <sup>216</sup>Prostate Cancer Canada, Toronto, Ontario, Canada. <sup>217</sup>National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. <sup>218</sup>Department of Biotechnology, Ministry of Science & Technology, Government of India, New Delhi, Delhi, India. <sup>219</sup>Science Writer, Garrett Park, MD, USA. <sup>220</sup>Cancer Research UK, London, UK. <sup>221</sup>Chinese Cancer Genome Consortium, Shenzhen, China. <sup>222</sup>Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. <sup>223</sup>Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. <sup>224</sup>National Cancer Center, Tokyo, Japan. <sup>225</sup>German Cancer Aid, Bonn, Germany. <sup>226</sup>Division of Cancer Genomics, National Cancer Center Research Institute, National Cancer Center, Tokyo, Japan. <sup>227</sup>Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan. <sup>228</sup>Japan

Agency for Medical Research and Development, Chiyoda-ku, Tokyo, Japan. <sup>229</sup>Medical Oncology, University and Hospital Trust of Verona, Verona, Italy. <sup>230</sup>University of Verona, Verona, Italy. <sup>231</sup>National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>232</sup>CAPHRI Research School, Maastricht University, Maastricht, The Netherlands. <sup>233</sup>Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>234</sup>University of California San Diego, San Diego, CA, USA. <sup>235</sup>PDXen Biosystems, Seoul, South Korea. <sup>236</sup>Electronics and Telecommunications Research Institute, Daejeon, South Korea. <sup>237</sup>Children's Hospital of Philadelphia, Philadelphia, PA, USA. <sup>238</sup>University of Melbourne Centre for Cancer Research, Melbourne, Victoria, Australia. <sup>239</sup>Syntekabio, Daejeon, South Korea. <sup>240</sup>AbbVie, North Chicago, IL, USA. <sup>241</sup>Genomics Research Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>242</sup>Department of Pediatric Immunology, Hematology and Oncology, University Hospital, Heidelberg, Germany. <sup>243</sup>Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM), Heidelberg, Germany. <sup>244</sup>Seven Bridges, Charlestown, MA, USA. <sup>245</sup>Health Sciences Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA. <sup>246</sup>Functional and Structural Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>247</sup>Leidos Biomedical Research, McLean, VA, USA. <sup>248</sup>CSRA Incorporated, Fairfax, VA, USA. <sup>249</sup>Department of Internal Medicine, Stanford University, Stanford, CA, USA. <sup>250</sup>Clinical Bioinformatics, Swiss Institute of Bioinformatics, Geneva, Switzerland. <sup>251</sup>Institute for Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Switzerland. <sup>252</sup>Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. <sup>253</sup>MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>254</sup>Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. <sup>255</sup>Office of Cancer Genomics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>256</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing, China. <sup>257</sup>Geneplus-Shenzhen, Shenzhen, China. <sup>258</sup>Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, MA, USA. <sup>259</sup>Department of Pediatrics, Harvard Medical School, Boston, MA, USA. <sup>260</sup>Technical University of Denmark, Lyngby, Denmark. <sup>261</sup>University of Copenhagen, Copenhagen, Denmark. <sup>262</sup>Department for BioMedical Research, University of Bern, Bern, Switzerland. <sup>263</sup>Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern, Switzerland. <sup>264</sup>Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland. <sup>265</sup>Department of Genitourinary Medical Oncology - Research, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>266</sup>Department of Urology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>267</sup>Korea Advanced Institute of Science and Technology, Daejeon, South Korea. <sup>268</sup>Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. <sup>269</sup>Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. <sup>270</sup>University of Milano Bicocca, Monza, Italy. <sup>271</sup>Sir Peter MacCallum Department of Oncology, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia. <sup>272</sup>Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX, USA. <sup>273</sup>Health Data Science Unit, University Clinics, Heidelberg, Germany. <sup>274</sup>Department for Biomedical Research, University of Bern, Bern, Switzerland. <sup>275</sup>Research Core Center, National Cancer Centre Korea, Goyang-si, South Korea. <sup>276</sup>Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland. <sup>277</sup>Harvard University, Cambridge, MA, USA. <sup>278</sup>Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>279</sup>Department of Information Technology, Ghent University, Ghent, Belgium. <sup>280</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. <sup>281</sup>Yale School of Medicine, Yale University, New Haven, CT, USA. <sup>282</sup>Division of Hematology-Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea. <sup>283</sup>Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul, South Korea. <sup>284</sup>Cheonan Industry-Academic Collaboration Foundation, Sangmyung University, Cheonan, South Korea. <sup>285</sup>Spanish National Cancer Research Centre, Madrid, Spain. <sup>286</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA. <sup>287</sup>Bern Center for Precision Medicine, University Hospital of Bern, University of Bern, Bern, Switzerland. <sup>288</sup>Englander Institute for Precision Medicine, Weill Cornell Medicine and New York Presbyterian Hospital, New York, NY, USA. <sup>289</sup>Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. <sup>290</sup>Pathology and Laboratory, Weill Cornell Medical College, New York, NY, USA. <sup>291</sup>cBio Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. <sup>292</sup>Department of Cell Biology, Harvard Medical School, Boston, MA, USA. <sup>293</sup>cBio Center, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>294</sup>CREST, Japan Science and Technology Agency, Tokyo, Japan. <sup>295</sup>Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo, Japan. <sup>296</sup>Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. <sup>297</sup>Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. <sup>298</sup>Department of Gene Technology, Tallinn University of Technology, Tallinn, Estonia. <sup>299</sup>Genetics & Genome Biology Program, SickKids Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>300</sup>Department of Information Technology, Ghent University, Interuniversitair Micro-Electronica Centrum (IMEC), Ghent, Belgium. <sup>301</sup>Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. <sup>302</sup>Oregon Health & Sciences University, Portland, OR, USA. <sup>303</sup>Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, Hong Kong, China. <sup>304</sup>The University of Texas Health Science Center at Houston, Houston, TX, USA. <sup>305</sup>Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA.

# Article

<sup>306</sup>The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH, USA. <sup>307</sup>The University of Texas School of Biomedical Informatics (SBMI) at Houston, Houston, TX, USA. <sup>308</sup>Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. <sup>309</sup>Physics Division, Optimization and Systems Biology Lab, Massachusetts General Hospital, Boston, MA, USA. <sup>310</sup>Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan. <sup>311</sup>Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, Germany. <sup>312</sup>Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany. <sup>313</sup>Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>314</sup>Computational Biology, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Jena, Germany. <sup>315</sup>Transcriptome Bioinformatics, LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany. <sup>316</sup>Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI, USA. <sup>317</sup>Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. <sup>318</sup>Research Center for Advanced Science and Technology, The University of Tokyo, Minato-ku, Tokyo, Japan. <sup>319</sup>Van Andel Research Institute, Grand Rapids, MI, USA. <sup>320</sup>Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>321</sup>Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>322</sup>The Hebrew University Faculty of Medicine, Jerusalem, Israel. <sup>323</sup>German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>324</sup>Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>325</sup>McKusick-Nathans Institute of Genetic Medicine, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>326</sup>Foundation Medicine, Cambridge, MA, USA. <sup>327</sup>Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. <sup>328</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>329</sup>University of Cambridge, Cambridge, UK. <sup>330</sup>Brandeis University, Waltham, MA, USA. <sup>331</sup>Hopp Children's Cancer Center (KITZ), Heidelberg, Germany. <sup>332</sup>Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>333</sup>A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia. <sup>334</sup>Oncology and Immunology, Dmitry Rogachev National Research Center of Pediatric Hematology, Moscow, Russia. <sup>335</sup>Skolkovo Institute of Science and Technology, Moscow, Russia. <sup>336</sup>Center for Medical Innovation, Seoul National University Hospital, Seoul, South Korea. <sup>337</sup>Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. <sup>338</sup>Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. <sup>339</sup>School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. <sup>340</sup>Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. <sup>341</sup>Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. <sup>342</sup>Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, USA. <sup>343</sup>Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL, USA. <sup>344</sup>Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, NY, USA. <sup>345</sup>Department of Bioengineering, and Department of Cellular and Molecular Medicine, Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. <sup>346</sup>Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore. <sup>347</sup>Department of Computer Science, University of Helsinki, Helsinki, Finland. <sup>348</sup>Institute of Biotechnology, University of Helsinki, Helsinki, Finland. <sup>349</sup>Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland. <sup>350</sup>Programme in Cancer & Stem Cell Biology, Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore. <sup>351</sup>Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK. <sup>352</sup>Department of Statistics, Columbia University, New York, NY, USA. <sup>353</sup>Duke-NUS Medical School, Singapore, Singapore. <sup>354</sup>School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. <sup>355</sup>The Kinghorn Cancer Centre, Cancer Division, Garvan Institute of Medical Research, University of New South Wales, Sydney, New South Wales, Australia. <sup>356</sup>MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Edinburgh, UK. <sup>357</sup>Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia. <sup>358</sup>Department of Bioinformatics, Division of Cancer Genomics, National Cancer Center Research Institute, National Cancer Center, Tokyo, Japan. <sup>359</sup>University of Glasgow, Glasgow, UK. <sup>360</sup>Academic Department of Medical Genetics, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. <sup>361</sup>MRC Cancer Unit, University of Cambridge, Cambridge, UK. <sup>362</sup>The University of Cambridge School of Clinical Medicine, Cambridge, UK. <sup>363</sup>MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. <sup>364</sup>Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Bearsden, UK. <sup>365</sup>School of Computing Science, University of Glasgow, Glasgow, UK. <sup>366</sup>South Western Sydney Clinical School, Faculty of Medicine, University of New South Wales, Liverpool, New South Wales, Australia. <sup>367</sup>West of Scotland Pancreatic Unit, Glasgow Royal Infirmary, Glasgow, UK. <sup>368</sup>University of Melbourne Centre for Cancer Research, Melbourne, Victoria, Australia. <sup>369</sup>Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. <sup>370</sup>Department of Surgery, University of Melbourne, Parkville, Victoria, Australia. <sup>371</sup>The Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia. <sup>372</sup>Walter + Eliza Hall Institute, Parkville, Victoria, Australia. <sup>373</sup>University of Cologne, Cologne, Germany. <sup>374</sup>The Edward S. Rogers Sr Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. <sup>375</sup>University of Ljubljana, Ljubljana, Slovenia. <sup>376</sup>Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA. <sup>377</sup>Research Institute, NorthShore University HealthSystem, Evanston, IL, USA. <sup>378</sup>Department

of Statistics, University of California Santa Cruz, Santa Cruz, CA, USA. <sup>379</sup>Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. <sup>380</sup>University of Toronto, Toronto, Ontario, Canada. <sup>381</sup>Department of Computer Science, Carleton College, Northfield, MN, USA. <sup>382</sup>Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. <sup>383</sup>Center for Psychiatric Genetics, NorthShore University HealthSystem, Evanston, IL, USA. <sup>384</sup>Argmix Consulting, North Vancouver, British Columbia, Canada. <sup>385</sup>Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>386</sup>Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>387</sup>The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>388</sup>Molecular and Medical Genetics, Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. <sup>389</sup>Department of Health Sciences, Faculty of Medical Sciences, Kyushu University, Fukuoka, Japan. <sup>390</sup>Baylor College of Medicine, Houston, TX, USA. <sup>391</sup>Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA. <sup>392</sup>Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany. <sup>393</sup>University Medical Center Hamburg-Eppendorf, Bioinformatics Core, Hamburg, Germany. <sup>394</sup>Earlham Institute, Norwich, UK. <sup>395</sup>Norwich Medical School, University of East Anglia, Norwich, UK. <sup>396</sup>The Institute of Cancer Research, London, UK. <sup>397</sup>University of East Anglia, Norwich, UK. <sup>398</sup>German Center for Infection Research (DZIF), Partner Site Hamburg-Borstel-Lübeck-Riems, Hamburg, Germany. <sup>399</sup>Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>400</sup>Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>401</sup>Victorian Institute of Forensic Medicine, Southbank, Victoria, Australia. <sup>402</sup>Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia. <sup>403</sup>University of Pennsylvania, Philadelphia, PA, USA. <sup>404</sup>Centre for Cancer Research, The Westmead Institute for Medical Research, Sydney, New South Wales, Australia. <sup>405</sup>Department of Gynaecological Oncology, Westmead Hospital, Sydney, New South Wales, Australia. <sup>406</sup>Genetics and Molecular Pathology, SA Pathology, Adelaide, South Australia, Australia. <sup>407</sup>Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. <sup>408</sup>Department of Gynaecological Oncology, Westmead Hospital, Sydney, New South Wales, Australia. <sup>409</sup>Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia. <sup>410</sup>Department of Clinical Pathology, University of Melbourne, Melbourne, Victoria, Australia. <sup>411</sup>Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. <sup>412</sup>Department of Gynaecological Oncology, Westmead Hospital, Sydney, New South Wales, Australia. <sup>413</sup>Westmead Clinical School, The Westmead Institute for Medical Research, Sydney, New South Wales, Australia. <sup>414</sup>Department of Surgery, Pancreas Institute, University and Hospital Trust of Verona, Verona, Italy. <sup>415</sup>Department of Surgery, Princess Alexandra Hospital, Brisbane, Queensland, Australia. <sup>416</sup>Surgical Oncology Group, Diamantina Institute, The University of Queensland, Brisbane, Queensland, Australia. <sup>417</sup>Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona, Italy. <sup>418</sup>ARC-Net Centre for Applied Research on Cancer, University and Hospital Trust of Verona, Verona, Italy. <sup>419</sup>Illawarra Shoalhaven Local Health District L3 Illawarra Cancer Care Centre, Wollongong Hospital, Wollongong, New South Wales, Australia. <sup>420</sup>Department of Pathology, University of Sydney, Sydney, New South Wales, Australia. <sup>421</sup>School of Biological Sciences, The University of Auckland, Auckland, New Zealand. <sup>422</sup>Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy. <sup>423</sup>Department of Medicine, Section of Endocrinology, University and Hospital Trust of Verona, Verona, Italy. <sup>424</sup>Department of Pathology, Queen Elizabeth University Hospital, Glasgow, UK. <sup>425</sup>Department of Medical Oncology, Beatson West of Scotland Cancer Centre, Glasgow, UK. <sup>426</sup>Academic Unit of Surgery, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow Royal Infirmary, Glasgow, UK. <sup>427</sup>Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia. <sup>428</sup>Discipline of Surgery, Western Sydney University, Penrith, New South Wales, Australia. <sup>429</sup>Institute of Cancer Sciences, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. <sup>430</sup>The Kinghorn Cancer Centre, Cancer Division, Garvan Institute of Medical Research, University of New South Wales, Sydney, New South Wales, Australia. <sup>431</sup>School of Environmental and Life Sciences, Faculty of Science, The University of Newcastle, Ourimbah, New South Wales, Australia. <sup>432</sup>Eastern Clinical School, Monash University, Melbourne, Victoria, Australia. <sup>433</sup>Epworth HealthCare, Richmond, Victoria, Australia. <sup>434</sup>Olivia Newton-John Cancer Research Institute, La Trobe University, Heidelberg, Victoria, Australia. <sup>435</sup>Melanoma Institute Australia, The University of Sydney, Wollstonecraft, New South Wales, Australia. <sup>436</sup>Children's Hospital at Westmead, The University of Sydney, Sydney, New South Wales, Australia. <sup>437</sup>Melanoma Institute Australia, The University of Sydney, Sydney, New South Wales, Australia. <sup>438</sup>Australian Institute of Tropical Health and Medicine, James Cook University, Douglas, Queensland, Australia. <sup>439</sup>Bioplatforms Australia, North Ryde, New South Wales, Australia. <sup>440</sup>Melanoma Institute Australia, Macquarie University, Wollstonecraft, New South Wales, Australia. <sup>441</sup>Children's Medical Research Institute, Sydney, New South Wales, Australia. <sup>442</sup>Melanoma Institute Australia, The University of Sydney, Wollstonecraft, New South Wales, Australia. <sup>443</sup>Centre for Cancer Research, The Westmead Millennium Institute for Medical Research, University of Sydney, Westmead Hospital, Sydney, New South Wales, Australia. <sup>444</sup>Translational Cancer Research Centre, The University of Sydney at the Westmead Institute, Sydney, New South Wales, Australia. <sup>445</sup>Discipline of Pathology, Sydney Medical School, The University of Sydney, Sydney, New South Wales, Australia. <sup>446</sup>School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia. <sup>447</sup>Melanoma Institute Australia, The University of Sydney, Wollstonecraft, New South Wales, Australia. <sup>448</sup>Royal Prince Alfred Hospital, Sydney, New South Wales, Australia. <sup>449</sup>Diagnostic Development, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>450</sup>Ontario

Tumour Bank, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>451</sup>PanCuRx Translational Research Initiative, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>452</sup>BioSpecimen Sciences Program, University Health Network, Toronto, Ontario, Canada. <sup>453</sup>Hepatobiliary/Pancreatic Surgical Oncology Program, University Health Network, Toronto, Ontario, Canada. <sup>454</sup>Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. <sup>455</sup>Division of Medical Oncology, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. <sup>456</sup>University of Nebraska Medical Center, Omaha, NE, USA. <sup>457</sup>BioSpecimen Sciences Program, University Health Network, Toronto, Ontario, Canada. <sup>458</sup>Transformative Pathology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>459</sup>University Health Network, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. <sup>460</sup>Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. <sup>461</sup>BioSpecimen Sciences, Laboratory Medicine (Toronto), Medical Biophysics, PanCuRX, Toronto, Ontario, Canada. <sup>462</sup>Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. <sup>463</sup>Department of Pathology, Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>464</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. <sup>465</sup>Department of Biochemistry and Molecular Medicine, University California at Davis, Sacramento, CA, USA. <sup>466</sup>Human Longevity, San Diego, CA, USA. <sup>467</sup>Department of Surgical Oncology, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. <sup>468</sup>Genome Informatics Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>469</sup>STTARR Innovation Facility, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. <sup>470</sup>Department of Pathology, Toronto General Hospital, Toronto, Ontario, Canada. <sup>471</sup>CRUK Manchester Institute and Centre, Manchester, UK. <sup>472</sup>Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. <sup>473</sup>Manchester Cancer Research Centre, Cancer Division, FBMH, University of Manchester, Manchester, UK. <sup>474</sup>Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. <sup>475</sup>Hefei University of Technology, Anhui, China. <sup>476</sup>State Key Laboratory of Cancer Biology and Xijing Hospital of Digestive Diseases, Fourth Military Medical University, Shaanxi, China. <sup>477</sup>Fourth Military Medical University, Shaanxi, China. <sup>478</sup>Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. <sup>479</sup>Department of Surgery, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China. <sup>480</sup>Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, UK. <sup>481</sup>Canadian Center for Computational Genomics, McGill University, Montreal, Quebec, Canada. <sup>482</sup>Department of Human Genetics, McGill University, Montreal, Quebec, Canada. <sup>483</sup>International Agency for Research on Cancer, Lyon, France. <sup>484</sup>McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada. <sup>485</sup>St James Institute of Oncology, University of Leeds, St James's University Hospital, Leeds, UK. <sup>486</sup>Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia. <sup>487</sup>Centre National de Génotypage, CEA - Institut de Génomique, Evry, France. <sup>488</sup>Department of Oncology, Gil Medical Center, Gachon University, Incheon, South Korea. <sup>489</sup>Department of Molecular Oncology, BC Cancer Agency, Vancouver, British Columbia, Canada. <sup>490</sup>Los Alamos National Laboratory, Los Alamos, NM, USA. <sup>491</sup>Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway. <sup>492</sup>Lund University, Lund, Sweden. <sup>493</sup>Translational Research Lab, Centre Léon Bérard, Lyon, France. <sup>494</sup>Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands. <sup>495</sup>Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. <sup>496</sup>Department of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. <sup>497</sup>Li Ka Shing Centre, Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>498</sup>Department of Oncology, University of Cambridge, Cambridge, UK. <sup>499</sup>Breast Cancer Translational Research Laboratory J. C. Heuson, Institut Jules Bordet, Brussels, Belgium. <sup>500</sup>Laboratory for Translational Breast Cancer Research, Department of Oncology, KU Leuven, Leuven, Belgium. <sup>501</sup>Translational Cancer Research Unit, GZA Hospitals St-Augustinus, Center for Oncological Research, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. <sup>502</sup>Department of Gynecology & Obstetrics and Department of Clinical Sciences, Skåne University Hospital, Lund University, Lund, Sweden. <sup>503</sup>Icelandic Cancer Registry, Icelandic Cancer Society, Reykjavik, Iceland. <sup>504</sup>Department of Medical Oncology, Josephine Nefkens Institute and Cancer Genomics Centre, Erasmus Medical Center, Rotterdam, The Netherlands. <sup>505</sup>National Genotyping Center, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. <sup>506</sup>Department of Pathology, Oslo University Hospital Ullevål, Oslo, Norway. <sup>507</sup>Faculty of Medicine and Institute of Clinical Medicine, University of Oslo, Oslo, Norway. <sup>508</sup>Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>509</sup>Department of Pathology, Skåne University Hospital, Lund University, Lund, Sweden. <sup>510</sup>Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands. <sup>511</sup>Department of Pathology, College of Medicine, Hanyang University, Seoul, South Korea. <sup>512</sup>Department of Pathology, Asan Medical Center, College of Medicine, Ulsan University, Songpa-gu, Seoul, South Korea. <sup>513</sup>The Netherlands Cancer Institute, Amsterdam, The Netherlands. <sup>514</sup>Department of Surgery, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Boston, MA, USA. <sup>515</sup>Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>516</sup>Department of Clinical Science, University of Bergen, Bergen, Norway. <sup>517</sup>Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>518</sup>The University of Queensland Centre for Clinical Research, The Royal Brisbane & Women's Hospital, Herston, Queensland, Australia. <sup>519</sup>Department of Pathology, Institut Jules Bordet, Brussels, Belgium. <sup>520</sup>Institute for Bioengineering and Biopharmaceutical Research (IBBR), Hanyang University, Seoul, South Korea. <sup>521</sup>University of Oslo, Oslo, Norway. <sup>522</sup>Institut Bergonié, Bordeaux, France. <sup>523</sup>Department of Research Oncology, Guy's Hospital, King's Health Partners AHSC, King's College London School of Medicine, London, UK. <sup>524</sup>University Hospital of Minjoo, INSERM UMR 1098, Besançon, France. <sup>525</sup>Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, UK. <sup>526</sup>East of Scotland Breast Service, Ninewells Hospital, Aberdeen, UK. <sup>527</sup>Oncologie Sénologie, ICM Institut Régional du Cancer, Montpellier, France. <sup>528</sup>Department of Radiation Oncology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. <sup>529</sup>University of Iceland, Reykjavik, Iceland. <sup>530</sup>Dundee Cancer Centre, Ninewells Hospital, Dundee, UK. <sup>531</sup>Institut Curie, INSERM Unit 830, Paris, France. <sup>532</sup>Department of Laboratory Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. <sup>533</sup>Department of General Surgery, Singapore General Hospital, Singapore, Singapore. <sup>534</sup>INCa-Synergie, Centre Léon Bérard, Université Lyon, Lyon, France. <sup>535</sup>Giovanni Paolo II/I.R.C.C.S. Cancer Institute, Bari, Italy. <sup>536</sup>Department of Biopathology, Centre Léon Bérard, Lyon, France. <sup>537</sup>Université Claude Bernard Lyon 1, Villeurbanne, France. <sup>538</sup>NCCS-VARI Translational Research Laboratory, National Cancer Centre Singapore, Singapore, Singapore. <sup>539</sup>Department of Pathology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands. <sup>540</sup>Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands. <sup>541</sup>Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany. <sup>542</sup>Institute of Human Genetics, University of Ulm, Ulm, Germany. <sup>543</sup>University Hospital of Ulm, Ulm, Germany. <sup>544</sup>Hematopathology Section, Institute of Pathology, Christian-Albrechts-University, Kiel, Germany. <sup>545</sup>Department of Human Genetics, Hannover Medical School, Hannover, Germany. <sup>546</sup>Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Düsseldorf, Germany. <sup>547</sup>Department of Internal Medicine/Hematology, Friedrich-Ebert-Hospital, Neumünster, Germany. <sup>548</sup>Pediatric Hematology and Oncology, University Hospital Muenster, Muenster, Germany. <sup>549</sup>Department of Pediatrics, University Hospital Schleswig-Holstein, Kiel, Germany. <sup>550</sup>Department of Medicine II, University of Würzburg, Würzburg, Germany. <sup>551</sup>Senckenberg Institute of Pathology, University of Frankfurt Medical School, Frankfurt, Germany. <sup>552</sup>Institute of Pathology, Charité—University Medicine Berlin, Berlin, Germany. <sup>553</sup>Department for Internal Medicine II, University Hospital Schleswig-Holstein, Kiel, Germany. <sup>554</sup>Institute for Medical Informatics Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. <sup>555</sup>Department of Hematology and Oncology, Georg-Augusts-University of Göttingen, Göttingen, Germany. <sup>556</sup>Institute of Cell Biology (Cancer Research), University of Duisburg-Essen, Essen, Germany. <sup>557</sup>MVZ Department of Oncology, PraxisClinic am Johannisplatz, Leipzig, Germany. <sup>558</sup>Institute of Pathology, Ulm University and University Hospital of Ulm, Ulm, Germany. <sup>559</sup>Department of Pathology, Robert-Bosch-Hospital, Stuttgart, Germany. <sup>560</sup>Pediatric Hematology and Oncology, University Hospital Giessen, Giessen, Germany. <sup>561</sup>Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. <sup>562</sup>Institute of Pathology, University of Wuerzburg, Wuerzburg, Germany. <sup>563</sup>Department of General Internal Medicine, University Kiel, Kiel, Germany. <sup>564</sup>Clinic for Hematology and Oncology, St-Antonius-Hospital, Eschweiler, Germany. <sup>565</sup>Department for Internal Medicine III, University of Ulm and University Hospital of Ulm, Ulm, Germany. <sup>566</sup>Neuroblastoma Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>567</sup>Department of Pediatric Oncology and Hematology, University of Cologne, Cologne, Germany. <sup>568</sup>University of Düsseldorf, Düsseldorf, Germany. <sup>569</sup>Department of Vertebrate Genomics/Otto Warburg Laboratory Gene Regulation and Systems Biology of Cancer, Max Planck Institute for Molecular Genetics, Berlin, Germany. <sup>570</sup>St Jude Children's Research Hospital, Memphis, TN, USA. <sup>571</sup>Heidelberg University Hospital, Heidelberg, Germany. <sup>572</sup>Genomics and Proteomics Core Facility High Throughput Sequencing Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>573</sup>Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>574</sup>University Medical Center Hamburg-Eppendorf, Hamburg, Germany. <sup>575</sup>Martini-Clinic, Prostate Cancer Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. <sup>576</sup>Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. <sup>577</sup>Division of Cancer Genome Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>578</sup>National Institute of Biomedical Genomics, Kalyani, India. <sup>579</sup>Advanced Centre for Treatment Research & Education in Cancer, Tata Memorial Centre, Navi Mumbai, India. <sup>580</sup>Department of Pathology, General Hospital of Treviso, Department of Medicine, University of Padua, Treviso, Italy. <sup>581</sup>Department of Medicine (DIMED), Surgical Pathology Unit, University of Padua, Padua, Italy. <sup>582</sup>Department of Hepatobiliary and Pancreatic Oncology, Hepatobiliary and Pancreatic Surgery Division, Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, Chuo-ku, Tokyo, Japan. <sup>583</sup>Department of Pathology, Keio University School of Medicine, Tokyo, Japan. <sup>584</sup>Department of Hepatobiliary and Pancreatic Oncology, National Cancer Center Hospital, Tokyo, Japan. <sup>585</sup>Department of Pathology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. <sup>586</sup>Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. <sup>587</sup>Gastric Surgery Division, Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, Tokyo, Japan. <sup>588</sup>Department of Gastroenterology and Hepatology, Yokohama City University Graduate School of Medicine, Kanagawa, Japan. <sup>589</sup>Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, University of Tokyo, Tokyo, Japan. <sup>590</sup>Department of Cancer Genome Informatics, Graduate School of Medicine, Osaka University, Osaka, Japan. <sup>591</sup>Hiroshima University, Hiroshima, Japan. <sup>592</sup>Tokyo Women's Medical University, Tokyo, Japan. <sup>593</sup>Osaka International Cancer Center, Osaka, Japan. <sup>594</sup>Wakayama Medical University, Wakayama, Japan. <sup>595</sup>Hokkaido University, Sapporo, Japan. <sup>596</sup>Division of Medical Oncology, National Cancer Centre, Singapore, Singapore. <sup>597</sup>Cholangiocarcinoma Screening and Care Program and Liver Fluke and Cholangiocarcinoma Research Centre, Faculty of Medicine, Khon Kaen University,

# Article

Khon Kaen, Thailand. <sup>598</sup>Lymphoma Genomic Translational Research Laboratory, National Cancer Centre, Singapore, Singapore. <sup>599</sup>Center of Digestive Diseases and Liver Transplantation, Fundeni Clinical Institute, Bucharest, Romania. <sup>600</sup>Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, School of Medicine, Keimyung University Dongsan Medical Center, Daegu, South Korea. <sup>601</sup>Pathology, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. <sup>602</sup>Hematology, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. <sup>603</sup>Department of Biochemistry and Molecular Biology, Faculty of Medicine, University Institute of Oncology-IUOPA, Oviedo, Spain. <sup>604</sup>Anatomia Patològica, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. <sup>605</sup>Spanish Ministry of Science and Innovation, Madrid, Spain. <sup>606</sup>Royal National Orthopaedic Hospital (Bolsover), London, UK. <sup>607</sup>Department of Pathology, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway. <sup>608</sup>Institute of Clinical Medicine and Institute of Oral Biology, University of Oslo, Oslo, Norway. <sup>609</sup>Research Department of Pathology, University College London Cancer Institute, London, UK. <sup>610</sup>East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. <sup>611</sup>Royal National Orthopaedic Hospital (Stanmore), London, UK. <sup>612</sup>Division of Orthopaedic Surgery, Oslo University Hospital, Oslo, Norway. <sup>613</sup>Department of Pathology (Research), University College London Cancer Institute, London, UK. <sup>614</sup>Radcliffe Department of Medicine, University of Oxford, Oxford, UK. <sup>615</sup>University of Pavia, Pavia, Italy. <sup>616</sup>Karolinska Institute, Stockholm, Sweden. <sup>617</sup>Wellcome Sanger Institute, Hinxton, UK. <sup>618</sup>University of Oxford, Oxford, UK. <sup>619</sup>Salford Royal NHS Foundation Trust, Salford, UK. <sup>620</sup>Gloucester Royal Hospital, Gloucester, UK. <sup>621</sup>Royal Stoke University Hospital, Stoke-on-Trent, UK. <sup>622</sup>St Thomas's Hospital, London, UK. <sup>623</sup>Imperial College NHS Trust, Imperial College London, London, UK. <sup>624</sup>Department of Histopathology, Salford Royal NHS Foundation Trust, Salford, UK. <sup>625</sup>Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. <sup>626</sup>Edinburgh Royal Infirmary, Edinburgh, UK. <sup>627</sup>Barking Havering and Redbridge University Hospitals NHS Trust, Romford, UK. <sup>628</sup>King's College London and Guy's and St Thomas' NHS Foundation Trust, London, UK. <sup>629</sup>Cambridge Oesophagogastric Centre, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. <sup>630</sup>Nottingham University Hospitals NHS Trust, Nottingham, UK. <sup>631</sup>St Luke's Cancer Centre, Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK. <sup>632</sup>University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>633</sup>Norfolk and Norwich University Hospital NHS Trust, Norwich, UK. <sup>634</sup>University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK. <sup>635</sup>University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. <sup>636</sup>Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK. <sup>637</sup>School of Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. <sup>638</sup>Wythenshawe Hospital, Manchester, UK. <sup>639</sup>Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK. <sup>640</sup>Royal Marsden NHS Foundation Trust, London and Sutton, London, UK. <sup>641</sup>University Hospital Southampton NHS Foundation Trust, Southampton, UK. <sup>642</sup>HCA Laboratories, London, UK. <sup>643</sup>University of Liverpool, Liverpool, UK. <sup>644</sup>Academic Urology Group, Department of Surgery, University of Cambridge, Cambridge, UK. <sup>645</sup>University of Oxford, Oxford, Oxford, UK. <sup>646</sup>Department of Urology, James Buchanan Brady Urological Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>647</sup>Second Military Medical University, Shanghai, China. <sup>648</sup>Department of Surgery and Cancer, Imperial College London, London, UK. <sup>649</sup>The Chinese University of Hong Kong, Shatin, Hong Kong, China. <sup>650</sup>Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of Oxford, Headington, Oxford, UK. <sup>651</sup>Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. <sup>652</sup>Department of Bioinformatics and Computational Biology and Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>653</sup>Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. <sup>654</sup>Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, British Columbia, Canada. <sup>655</sup>Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>656</sup>University Health Network, Toronto, Ontario, Canada. <sup>657</sup>Department of Pathology and Laboratory Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>658</sup>Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA. <sup>659</sup>Research Health Analytics and Informatics, University Hospitals Cleveland Medical Center, Cleveland, OH, USA. <sup>660</sup>Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, Alberta, Canada. <sup>661</sup>Department of Surgery and Department of Oncology, University of Calgary, Calgary, Alberta, Canada. <sup>662</sup>Buck Institute for Research on Aging, Novato, CA, USA. <sup>663</sup>Duke University Medical Center, Durham, NC, USA. <sup>664</sup>USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA. <sup>665</sup>The Preston Robert Tisch Brain Tumor Center, Duke University Medical Center, Durham, NC, USA. <sup>666</sup>Department of Dermatology and Department of Pathology, Yale University, New Haven, CT, USA. <sup>667</sup>Fox Chase Cancer Center, Philadelphia, PA, USA. <sup>668</sup>Department of Surgery, Division of Thoracic Surgery, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>669</sup>University of Michigan Comprehensive Cancer Center, Ann Arbor, MI, USA. <sup>670</sup>University of Alabama at Birmingham, Birmingham, AL, USA. <sup>671</sup>Division of Anatomic Pathology, Mayo Clinic, Rochester, MN, USA. <sup>672</sup>Division of Experimental Pathology, Mayo Clinic, Rochester, MN, USA. <sup>673</sup>Department of Oncology, The Johns Hopkins School of Medicine, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, MD, USA. <sup>674</sup>International Genomics Consortium, Phoenix, AZ, USA. <sup>675</sup>Department of Pediatrics and Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>676</sup>Department of Pathology, UPMC Shadyside, Pittsburgh, PA, USA. <sup>677</sup>Center for Cancer Genomics, National

Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>678</sup>Department of Neuro-Oncology, Istituto Neurologico Besta, Milan, Italy. <sup>679</sup>University of Queensland Thoracic Research Centre, The Prince Charles Hospital, Brisbane, Queensland, Australia. <sup>680</sup>Department of Neurosurgery, University of Florida, Gainesville, FL, USA. <sup>681</sup>Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>682</sup>Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>683</sup>Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>684</sup>Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA. <sup>685</sup>Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, USA. <sup>686</sup>Department of Internal Medicine, Division of Medical Oncology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>687</sup>University of Tennessee Health Science Center for Cancer Research, Memphis, TN, USA. <sup>688</sup>Centre for Translational and Applied Genomics, British Columbia Cancer Agency, Vancouver, British Columbia, Canada. <sup>689</sup>Department of Pathology & Immunology, Baylor College of Medicine, Houston, TX, USA. <sup>690</sup>Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX, USA. <sup>691</sup>Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>692</sup>Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. <sup>693</sup>Indivumed, Hamburg, Germany. <sup>694</sup>Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, School of Medicine, Keimyung University Dong-san Medical Center, Daegu, South Korea. <sup>695</sup>Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>696</sup>Department of Surgery, School of Medicine and Health Science, The George Washington University, Washington, DC, USA. <sup>697</sup>Endocrine Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>698</sup>National Cancer Center, Gyeonggi, South Korea. <sup>699</sup>LSBio, LLC Biobank, Chestertown, MD, USA. <sup>700</sup>Gynecologic Oncology, NYU Laura and Isaac Perlmutter Cancer Center, New York University, New York, NY, USA. <sup>701</sup>Division of Oncology, Stem Cell Biology Section, Washington University School of Medicine, St Louis, MO, USA. <sup>702</sup>Urologic Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>703</sup>Institute for Systems Biology, Seattle, WA, USA. <sup>704</sup>Center for Personalized Medicine, Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, CA, USA. <sup>705</sup>Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. <sup>706</sup>Department of Surgery, Duke University, Durham, NC, USA. <sup>707</sup>Department of Obstetrics, Gynecology and Reproductive Services, University of California San Francisco, San Francisco, CA, USA. <sup>708</sup>Department of Neurology and Department of Neurosurgery, Henry Ford Hospital, Detroit, MI, USA. <sup>709</sup>Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. <sup>710</sup>Department of Pathology, Roswell Park Cancer Institute, Buffalo, NY, USA. <sup>711</sup>Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, Washington University School of Medicine, St Louis, MO, USA. <sup>712</sup>Department of Palliative, Rehabilitation and Integrative Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>713</sup>Penrose St Francis Health Services, Colorado Springs, CO, USA. <sup>714</sup>The University of Chicago, Chicago, IL, USA. <sup>715</sup>Department of Neurology, Mayo Clinic, Rochester, MN, USA. <sup>716</sup>Center for Liver Cancer, Research Institute and Hospital, National Cancer Center, Gyeonggi, South Korea. <sup>717</sup>Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>718</sup>NYU Langone Medical Center, New York, NY, USA. <sup>719</sup>Department of Hematology and Medical Oncology, Cleveland Clinic, Cleveland, OH, USA. <sup>720</sup>Department of Genetics, Department of Pathology and Laboratory Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>721</sup>Helen F. Graham Cancer Center at Christiana Care Health Systems, Newark, DE, USA. <sup>722</sup>Cureline, South San Francisco, CA, USA. <sup>723</sup>Department of Obstetrics and Gynecology, Medical College of Wisconsin, Milwaukee, WI, USA. <sup>724</sup>Hematology and Medical Oncology, Winship Cancer Institute of Emory University, Atlanta, GA, USA. <sup>725</sup>Vanderbilt Ingram Cancer Center, Vanderbilt University, Nashville, TN, USA. <sup>726</sup>Ohio State University College of Medicine and Arthur G. James Comprehensive Cancer Center, Columbus, OH, USA. <sup>727</sup>Research Computing Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>728</sup>Analytical Biological Services, Wilmington, DE, USA. <sup>729</sup>Department of Dermatology, University Hospital Essen, Westdeutsches Tumorzentrum and German Cancer Consortium, Essen, Germany. <sup>730</sup>University of Pittsburgh, Pittsburgh, PA, USA. <sup>731</sup>Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, MD, USA. <sup>732</sup>Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. <sup>733</sup>Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>734</sup>Department of Gynecologic Oncology and Reproductive Medicine, and Center for RNA Interference and Non-Coding RNA, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>735</sup>Department of Urology, Mayo Clinic, Rochester, MN, USA. <sup>736</sup>Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>737</sup>Department of Neurosurgery, Department of Hematology and Department of Medical Oncology, Winship Cancer Institute and School of Medicine, Emory University, Atlanta, GA, USA. <sup>738</sup>Georgia Regents University Cancer Center, Augusta, GA, USA. <sup>739</sup>Thoracic Oncology Laboratory, Mayo Clinic, Rochester, MN, USA. <sup>740</sup>Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. <sup>741</sup>Department of Obstetrics & Gynecology, Division of Gynecologic Oncology, Mayo Clinic, Rochester, MN, USA. <sup>742</sup>International Institute for Molecular Oncology, Poznań, Poland. <sup>743</sup>Poznan University of Medical Sciences, Poznań, Poland. <sup>744</sup>Edison Family Center for Genome Sciences and Systems Biology, Washington University, St Louis, MO, USA. <sup>745</sup>These authors jointly supervised this work: Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Lincoln D. Stein. \*e-mail: pc8@sanger.ac.uk; gadgetz@broadinstitute.org; korbel@embl.de; jstuart@ucsc.edu; lincoln.stein@gmail.com



## Methods

### Samples

We compiled an inventory of matched tumour–normal whole-cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, although a small number of donors had multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (1) matched tumour and normal specimen pair; (2) a minimal set of clinical fields; and (3) characterization of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads.

We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014 (Extended Data Table 1). After quality assurance (Supplementary Methods 2.5), data from 176 donors were excluded as unusable, 75 had minor issues that could affect some analyses (grey-listed donors) and 2,583 had data of optimal quality (white-listed donors) (Supplementary Table 1). Across the 2,658 white- and grey-listed donors, whole-genome sequences were available from 2,605 primary tumours and 173 metastases or local recurrences. Matching normal samples were obtained from blood (2,064 donors), tissue adjacent to the primary tumour (87 donors) or from distant sites (507 donors). Whole-genome sequencing data were available for tumour and normal DNA for the entire cohort. The mean read coverage was 39× for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38× and 60× (Supplementary Fig. 1). The majority of specimens (65.3%) were sequenced using 101-bp paired-end reads. An additional 28% were sequenced with 100-bp paired-end reads. Of the remaining specimens, 4.7% were sequenced with read lengths longer than 101 bp, and 1.9% with read lengths shorter than 100 bp. The distribution of read lengths by tumour cohort is shown in Supplementary Fig. 11. Median read length for whole-genome sequencing paired-end reads was 101 bp (mean = 106.2, s.d. = 16.7; minimum–maximum = 50–151). RNA-sequencing data were collected and re-analysed centrally for 1,222 donors, including 1,178 primary tumours, 67 metastases or local recurrences and 153 matched normal tissue samples adjacent to the primary tumour.

Demographically, the cohort included 1,469 men (55%) and 1,189 women (45%), with a mean age of 56 years (range, 1–90 years) (Supplementary Table 1). Using population ancestry-differentiated single nucleotide polymorphisms, the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects (Supplementary Table 1).

We consolidated histopathology descriptions of the tumour samples, using the ICD-O-3 tumour site controlled vocabulary<sup>89</sup>. Overall, the PCAWG dataset comprises 38 distinct tumour types (Extended Data Table 1 and Supplementary Table 1). Although the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely owing to differences among contributing ICGC/TCGA groups in the numbers of sequenced samples.

### Uniform processing and somatic variant calling

To generate a consistent set of somatic mutation calls that could be used for cross-tumour analyses, we analysed all 6,835 samples using a uniform set of algorithms for alignment, variant calling and quality control (Extended Data Fig. 1, Supplementary Fig. 2, Supplementary Table 3 and Supplementary Methods 2). We used the BWA-MEM algorithm<sup>90</sup> to align each tumour and normal sample to human reference build hs37d5 (as used in the 1000 Genomes Project<sup>91</sup>). Somatic mutations were identified in the aligned data using three established pipelines, which were run independently on each tumour–normal pair. Each of the three pipelines—labelled ‘Sanger’<sup>92–95</sup>, ‘EMBL/DKFZ’<sup>96,97</sup> and ‘Broad’<sup>98–101</sup> after the computational biology groups that created or assembled

them—consisted of multiple software packages for calling somatic SNVs, small indels, CNAs and somatic SVs (with intrachromosomal SVs defined as those >100 bp). Two additional variant algorithms<sup>102,103</sup> were included to further improve accuracy across a broad range of clonal and subclonal mutations. We tested different merging strategies using validation data, and chose the optimal method for each variant type to generate a final consensus set of mutation calls (Supplementary Methods S2.4).

Somatic retrotransposition events, including Alu and LINE-1 insertions<sup>72</sup>, L1-mediated transductions<sup>73</sup> and pseudogene formation<sup>104</sup>, were called using a dedicated pipeline<sup>73</sup>. We removed these retrotransposition events from the somatic SV call-set. Mitochondrial DNA mutations were called using a published algorithm<sup>105</sup>. RNA-sequencing data were uniformly processed to quantify normalized gene-level expression, splicing variation and allele-specific expression, and to identify fusion transcripts, alternative promoter usage and sites of RNA editing<sup>8</sup>.

### Integration, phasing and validation of germline variant call-sets

Calls of common ( $\geq 1\%$  frequency in PCAWG) and rare ( $< 1\%$ ) germline variants including single-nucleotide polymorphisms, indels, SVs and mobile-element insertions (MEIs) were generated using a population-scale genetic polymorphism-detection approach<sup>91,106</sup>. The uniform germline data-processing workflow comprised variant identification using six different variant-calling algorithms<sup>96,107,108</sup> and was orchestrated using the Butler workflow system<sup>109</sup>.

We performed call-set benchmarking, merging, variant genotyping and statistical haplotype-block phasing<sup>91</sup> (Supplementary Methods 3.4). Using this strategy, we identified 80.1 million germline single-nucleotide polymorphisms, 5.9 million germline indels, 1.8 million multi-allelic short ( $< 50$  bp) germline variants, as well as germline SVs  $\geq 50$  bp in size including 29,492 biallelic deletions and 27,254 MEIs (Supplementary Table 2). We statistically phased this germline variant set using haplotypes from the 1000 Genomes Project<sup>91</sup> as a reference panel, yielding an N50-phased block length of 265 kb based on haploid chromosomes from donor-matched tumour genomes. Precision estimates for germline SNVs and indels were  $> 99\%$  for the phased merged call-set, and sensitivity estimates ranged from 92% to 98%.

### Core alignment and variant calling by cloud computing

The requirement to uniformly realign and call variants on nearly 5,800 whole genomes (tumour plus normal) presented considerable computational challenges, and raised ethical issues owing to the use of data from different jurisdictions (Extended Data Table 2). To process the data, we adopted a cloud-computing architecture<sup>26</sup> in which the alignment and variant calling was spread across 13 data centres on 3 continents, representing a mixture of commercial, infrastructure-as-a-service, academic cloud compute and traditional academic high-performance computer clusters (Supplementary Table 3). Together, the effort used 10 million CPU-core hours.

To generate reproducible variant calling across the 13 data centres, we built the core pipelines into Docker containers<sup>28</sup>, in which the workflow description, required code and all associated dependencies were packaged together in stand-alone packages. These heavily tested, extensively validated workflows are available for download (Box 1).

### Validation, benchmarking and merging of somatic variant calls

To evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep-sequencing validation experiment (Supplementary Notes 1). We selected a pilot set of 63 representative tumour–normal pairs, on which we ran the 3 core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the PCAWG SNV Calling Methods Working Group. Sufficient DNA remained for 50 of the 63 cases for validation, which was performed by hybridization of tumour and matched normal DNA to a custom RNA bait set, followed

# Article

by deep sequencing, as previously described<sup>29</sup>. Although performed using the same sequencing chemistry as the original whole-genome sequencing analyses, the considerably greater depth achieved in the validation experiment enabled accurate assessment of sensitivity and precision of variant calls. Variant calls in repeat-masked regions were not tested, owing to the challenge of designing reliable validation probes in these areas.

The 3 core pipelines had individual estimates of sensitivity of 80–90% to detect a true somatic SNV called by any of the 13 pipelines; with >95% of SNV calls made by each of the core pipelines being genuine somatic variants (Fig. 1a). For indels—a more-challenging class of variants to identify in short-read sequencing data—the 3 core algorithms had individual sensitivity estimates in the range of 40–50%, with precision 70–95% (Fig. 1b). Validation of SV calls is inherently more difficult, as methods based on PCR or hybridization to RNA baits often fail to isolate DNA that spans the breakpoint. To assess the accuracy of SV calls, we therefore used the property that an SV must either generate a copy-number change or be balanced, whereas artefactual calls will not respect this property. For individual SV-calling algorithms, we estimated precision to be in the range of 80–95% for samples in the 63-sample pilot dataset.

Next, we examined multiple methods for merging calls made by several algorithms into a single definitive call-set to be used for downstream analysis. The final consensus calls for SNVs were based on a simple approach that required two or more methods to agree on a call. For indels, because methods were less concordant, we used stacked logistic regression<sup>110,111</sup> to integrate the calls. The merged SV set includes all calls made by two or more of the four primary SV-calling algorithms<sup>96,100,112,113</sup>. Consensus CNA calls were obtained by joining the outputs of six individual CNA-calling algorithms with SV consensus breakpoints to obtain base-pair resolution CNAs (Supplementary Methods 2.4.3). Consensus purity and ploidy were derived, and a multitier system was developed for consensus copy-number calls (Supplementary Methods 2.4.3, and described in detail elsewhere<sup>7</sup>).

Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (90% confidence interval, 88–98%) and 95% (90% confidence interval, 71–99%), respectively, for SNVs (Extended Data Fig. 2). For somatic indels, sensitivity and precision were 60% (90% confidence interval, 34–72%) and 91% (90% confidence interval, 73–96%), respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one calling pipeline; precision was estimated to be 97.5%. That is, 97.5% of SVs in the merged SV call-set had an associated copy-number change or balanced partner rearrangement. The improvement in calling accuracy from combining different pipelines was most noticeable in variants that had low variant allele fractions, which are likely to originate from subclonal populations of the tumour (Fig. 1c, d). There remains much work to be done to improve indel calling software; we still lack sensitivity for calling even fully clonal complex indels from short-read sequencing data.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The PCAWG-generated alignments, somatic variant calls, annotations and derived datasets are available for general research use for browsing and download at <http://dcc.icgc.org/pcawg/> (Box 1 and Supplementary Table 4). In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identifying information, such as germline alleles and underlying read data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP ([https://dbgap.ncbi.nlm.nih.gov/aa/wga](https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login)

[cgi?page=login](https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login)) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Beyond the core sequence data and variant call-sets, the analyses in this paper used a number of datasets that were derived from the variant calls (Supplementary Table 4). The individual datasets are available at Synapse (<https://www.synapse.org/>), and are denoted with synXXXXX accession numbers; all these datasets are also mirrored at <https://dcc.icgc.org>, with full links, filenames, accession numbers and descriptions detailed in Supplementary Table 4. The datasets encompass: clinical data from each patient including demographics, tumour stage and vital status (syn10389158); harmonized tumour histopathology annotations using a standardised hierarchical ontology (syn1038916); inferred purity and ploidy values for each tumour sample (syn8272483); driver mutations for each patient from their cancer genome spanning all classes of variant, and coding versus non-coding drivers (syn11639581); mutational signatures inferred from PCAWG donors (syn11804065), including APOBEC mutagenesis (syn7437313); and transcriptional data from RNA sequencing, including gene expression levels (syn5553985, syn5553991, syn8105922) and gene fusions (syn10003873, syn7221157).

## Code availability

Computational pipelines for calling somatic mutations are available to the public at <https://dockstore.org/organizations/PCAWG/collections/PCAWG>. A range of data-visualization and -exploration tools are also available for the PCAWG data (Box 1).

89. NCI SEER. *ICD-O-3 Coding Materials* (2018).
90. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
91. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
92. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinformatics* **56**, 15.9.1–15.9.17 (2016).
93. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
94. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
95. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
96. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
97. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
98. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
99. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
100. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
101. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
102. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).
103. Fan, Y. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
104. Cooke, S. L. et al. Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **5**, 3644 (2014).
105. Ju, Y. S. et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, e02935 (2014).
106. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
107. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://arxiv.org/abs/1207.3907> (2012).
108. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

109. Yakneen, S., Waszak, S. M., Gertz, M. & Korbel, J. O. & PCAWG Consortium. Butler enables rapid cloud-based analysis of thousands of human genomes. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-019-0360-3> (2020).
110. Kim, S. Y., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics* **15**, 154 (2014).
111. Breiman, L. Stacked regressions. *Mach. Learn.* **24**, 49–64 (1996).
112. Campbell, P. J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
113. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).

**Acknowledgements** We thank research participants who donated samples and data, the physicians and clinical staff who contributed to sample annotation and collection, and the numerous funding agencies that contributed to the collection and analysis of this dataset.

**Author contributions** Writing committee leads: Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein. Head of project management: Jennifer L. Jennings. Sample collection: major contributions from Marc D. Perry, Hardeep K. Nahal-Bose; led by B. F. Francis Ouellette. Histopathology harmonization: major contribution from Constance H. Li; further contributions from Esther Rheinbay, G. Petur Nielsen, Dennis C. Sgroi, Chin-Lee Wu, William C. Faquin, Vikram Deshpande, Paul C. Boutros, Alexander J. Lazar, Katherine A. Hoadley; led by Lincoln D. Stein, David N. Louis. Uniform processing, somatic, germline variant calling: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Junjun Zhang, Wenyi Wang, David A. Wheeler; led by Li Ding, Jared T. Simpson. Core alignment, variant calling by cloud computing: major contributions from Christina K. Yung, Brian D. O'Connor, Sergei Yakneen, Junjun Zhang; further contributions from Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M. Raine, Adam P. Butler, Romina Royo, Gordon Saksena, Matthias Schlesner, Solomon I. Shorser, Miguel Vazquez. Integration, phasing, validation of germline variant callsets: major contributions from Tobias Rausch, Grace Tiao, Sebastian M. Waszak, Bernardo Rodriguez-Martin, Suyash Shringarpure, Dai-Ying Wu; further contributions from Sergei Yakneen, German M. Demidov, Olivier Delaneau, Shuto Hayashi, Seiya Imoto, Nina Habermann, Ayellet V. Segre, Erik Garrison, Andy Cafferkey, Eva G. Alvarez, José María Heredia-Genestar, Francesc Muyas, Oliver Drechsel, Alicia L. Bruzos, Javier Temes, Jorge Zamora, L. Jonathan Dursi, Adrian Baez-Ortega, Hyung-Lae Kim, Matthew H. Bailey, R. Jay Mashl, Kai Ye, Ivo Buchhalter, Anthony DiBiase, Kuan-lin Huang, Ivica Letunic, Michael D. McLellan, Steven J. Newhouse, Matthias Schlesner, Tal Shmaya, Sushant Kumar, David C. Wedge, Mark H. Wright, Venkata D. Yellapantula, Mark Gerstein, Ekta Khurana, Tomas Marques-Bonet, Arcadi Navarro, Carlos D. Bustamante, Jared T. Simpson, Li Ding, Reiner Siebert, Hidewaki Nakagawa, Douglas F. Easton; led by Stephan Ossowski, Jose M. C. Tubio, Gad Getz, Francisco M. De La Vega, Xavier Estivill, Jan O. Korbel. Validation, benchmarking, merging of somatic variant calls: major contribution from L. Jonathan Dursi; further contributions from David A. Wheeler, Christina K. Yung; led by Li Ding, Jared T. Simpson. Data and code availability: major contribution from Junjun Zhang; further contributions from Christina K. Yung, Sergei Yakneen, Denis Yuen, George L. Mihaescu, Larsson Omberg; led by Vincent Ferretti. Pan-cancer burden of somatic mutations: major contribution from Junjun Zhang; led by Peter J. Campbell. Panorama of driver mutations in human cancer: led by Radhakrishnan Sabarinathan, Oriol Pich, Abel Gonzalez-Perez. PCAWG tumours with no apparent driver mutations: major contribution from Esther Rheinbay; further contributions from Amaro Taylor-Weiner, Radhakrishnan Sabarinathan; led by Peter J. Campbell, Gad Getz. Patterns, oncogenicity of kataegis, chromoplexy: major contributions from Matthew W. Fittall, Jonas Demeulemeester, Maxime Tarabichi; further contributions from Nicola D. Roberts, Peter J. Campbell, Jan O. Korbel; led by Peter Van Loo. Patterns, oncogenicity of chromothripsis: major contributions from Maxime Tarabichi, Jonas Demeulemeester, Matthew W. Fittall; further contributions from Isidro Cortes-Ciriano, Lara Urban, Peter J. Park, Peter J. Campbell, Jan O. Korbel; led by Peter Van Loo. Timing-clustered mutational processes during tumour evolution: major contributions from Jonas Demeulemeester, Maxime Tarabichi, Matthew W. Fittall; further contributions from Jan O. Korbel, Peter J. Campbell; led by Peter Van Loo. Germline effects on somatic mutation: major contributions from Sebastian M. Waszak, Bin Zhu, Bernardo Rodriguez-Martin, Esa Pitkanen, Tobias Rausch; further contributions from Yilong Li, Natalie Saini, Leszek J. Klimczak, Joachim Weischenfeldt, Nikos Sidiropoulos, Ludmil B. Alexandrov, Francesc Muyas, Raquel Rabionet, Georgia Escaramis, Adrian Baez-Ortega, Mattia Bosio, Aliaksei Z. Holik, Hana Susak, Eva G. Alvarez, Alicia L. Bruzos, Javier Temes, Aparna Prasad, Nina Habermann, Serap Erkek, Lara Urban, Claudia Calabrese, Benjamin Raeder, Eoghan Harrington, Simon Mayes, Daniel Turner, Sissel Juul, Steven A. Roberts, Lei Song, Roelof Koster, Lisa Mirabello, Xing Hua, Tomas J. Tanskanen, Marta Tojo, David C. Wedge, Jorge Zamora, Jieming Chen, Lauri A. Aaltonen, Gunnar Ratsch, Roland F. Schwarz, Atul J. Butte, Alvis Brazma, Peter J. Campbell, Stephen J. Chanock, Nilanjan Chatterjee, Oliver Stegle, Olivier Harismendy; led by G. Steven Bova, Dmitry A. Gordenin, Jose M. C. Tubio, Douglas F. Easton, Xavier Estivill, Jan O. Korbel. Replicative immortality: major contribution from David Haan; further contributions from Lina Sieverling, Lars Feuerbach; led by Lincoln D. Stein, Joshua M. Stuart. Ethical considerations of genomic cloud computing: led by Don Chalmers, Yann Joly, Bartha Knoppers, Fruzsina Molnar-Gabor, Jan O. Korbel, Mark Phillips, Adrian Thorogood, David Townsend. Online resources for data access, visualization, exploration and analysis: major contributions from Mary Goldman, Junjun Zhang, Nuno A. Fonseca; further contributions from Qian Xiang, Brian Craft, Elena Pineiro-Yanez, Alfonso Munoz, Robert Petryszak, Anja Fullgrabe, Fatima Al-Shahrour, Maria Keays, David Haussler, John Weinstein, Wolfgang Huber, Alfonso Valencia, Irene Papatheodorou, Jingchun Zhu; led by Brian D. O'Connor, Lincoln D. Stein, Alvis Brazma, Vincent Ferretti, Miguel Vazquez. The 63-sample pilot-analysis validation process: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Yu Fan, David Torrents, Matthias Bieg, Paul C. Boutros, Ken Chen, Zechen Chong, Kristian Cibulskis, Oliver Drechsel, Roland Eils, Robert S. Fulton, Josep Gelpi, Mark Gerstein, Santiago Gonzalez, Gad Getz, Ivo G. Gut, Faraz Hach, Michael Heindol, Taobo Hu, Vincent Huang, Barbara Hutter, Hyung-Lae Kim, Natalie Jager, Jongsun Jung, Sushant Kumar,

Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Ivica Letunic, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, R. Jay Mashl, Michael D. McLellan, Ana Milovanovic, Morten Muhlig Nielsen, Brian D. O'Connor, Stephan Ossowski, Nagarajan Paramasivam, Jakob Skou Pedersen, Marc D. Perry, Montserrat Puiggros, Romina Royo, Esther Rheinbay, S. Cenik Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Grace Tiao, Jeremiah A. Wala, Jiayin Wang, Wenyi Wang, Sebastian M. Waszak, Joachim Weischenfeldt, Michael Wendt, Johannes Werner, Zhenggang Wu, Hong Xue, Sergei Yakneen, Takafumi N. Yamaguchi, Kai Ye, Venkata Yellapantula, Junjun Zhang, David A. Wheeler; led by Li Ding, Jared T. Simpson. Processing of validation data: major contributions from Christina K. Yung, Brian D. O'Connor, Sergei Yakneen, Junjun Zhang; further contributions from Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M. Raine, Romina Royo, Gordon Saksena, Matthias Schlesner, Solomon I. Shorser, Miguel Vazquez, Joachim Weischenfeldt, Denis Yuen, Adam P. Butler, Brandi N. Davis-Dusenbery, Roland Eils, Vincent Ferretti, Robert L. Grossman, Olivier Harismendy, Youngwook Kim, Hidewaki Nakagawa, Steven J. Newhouse, David Torrents; led by Lincoln D. Stein. Whole-genome sequencing somatic variant calling: major contribution from Junjun Zhang; further contributions from Christina K. Yung, Solomon I. Shorser. Whole-genome alignment: Keiran M. Raine, Junjun Zhang, Brian D. O'Connor. DKFZ pipeline: Kortine Kleinheinz, Tobias Rausch, Jan O. Korbel, Ivo Buchhalter, Michael C. Heindol, Barbara Hutter, Natalie Jager, Nagarajan Paramasivam, Matthias Schlesner. EMBL pipeline: Joachim Weischenfeldt. Sanger pipeline: Keiran M. Raine, Jonathan Hinton, David R. Jones, Andrew Menzies, Lucy Stebbings, Adam P. Butler. Broad pipeline: Gordon Saksena, Dimitri Livitz, Esther Rheinbay, Julian M. Hess, Ignaty Leshchiner, Chip Stewart, Grace Tiao, Jeremiah A. Wala, Amaro Taylor-Weiner, Mara Rosenberg, Andrew J. Dunford, Manasvi Gupta, Marcin Imielinski, Matthew Meyerson, Rameen Beroukhim, Gad Getz. MuSE Pipeline: Yu Fan, Wenyi Wang. Consensus somatic SNV/indel annotation: Andrew Menzies, Matthias Schlesner, Juri Reimand, Priyanka Dhingra, Ekta Khurana. Somatic SNV, indel merging: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Yu Fan, David Torrents, Matthias Bieg, Paul C. Boutros, Ken Chen, Zechen Chong, Kristian Cibulskis, Oliver Drechsel, Roland Eils, Robert S. Fulton, Josep Gelpi, Mark Gerstein, Santiago Gonzalez, Gad Getz, Ivo G. Gut, Faraz Hach, Michael Heindol, Taobo Hu, Vincent Huang, Barbara Hutter, Hyung-Lae Kim, Natalie Jager, Jongsun Jung, Sushant Kumar, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Ivica Letunic, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, R. Jay Mashl, Michael D. McLellan, Ana Milovanovic, Morten Muhlig Nielsen, Brian D. O'Connor, Stephan Ossowski, Nagarajan Paramasivam, Jakob Skou Pedersen, Marc D. Perry, Montserrat Puiggros, Romina Royo, Esther Rheinbay, S. Cenik Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Grace Tiao, Jeremiah A. Wala, Jiayin Wang, Wenyi Wang, Sebastian M. Waszak, Joachim Weischenfeldt, Michael Wendt, Johannes Werner, Zhenggang Wu, Hong Xue, Sergei Yakneen, Takafumi N. Yamaguchi, Kai Ye, Venkata Yellapantula, Junjun Zhang, David A. Wheeler; major contributions from Li Ding, Jared T. Simpson. Somatic SV merging: Joachim Weischenfeldt, Francesco Favero, Yilong Li. Somatic CNA merging: Stefan Dentre, Jeff Wintersinger, Ignaty Leshchiner. Oxidative artefact filtration: Dimitri Livitz, Ignaty Leshchiner, Chip Stewart, Esther Rheinbay, Gordon Saksena, Gad Getz. Strand bias filtration: Matthias Bieg, Ivo Buchhalter, Johannes Werner, Matthias Schlesner. miniBAM generation: Jeremiah A. Wala, Gordon Saksena, Rameen Beroukhim, Gad Getz. Germline variant identification from whole-genome sequencing: major contributions from Tobias Rausch, Grace Tiao, Sebastian M. Waszak, Bernardo Rodriguez-Martin, Suyash Shringarpure, Dai-Ying Wu; further contributions from Sergei Yakneen, German M. Demidov, Olivier Delaneau, Shuto Hayashi, Seiya Imoto, Nina Habermann, Ayellet V. Segre, Erik Garrison, Andy Cafferkey, Eva G. Alvarez, Alicia L. Bruzos, Jorge Zamora, José María Heredia-Genestar, Francesc Muyas, Oliver Drechsel, L. Jonathan Dursi, Adrian Baez-Ortega, Hyung-Lae Kim, Matthew H. Bailey, R. Jay Mashl, Kai Ye, Ivo Buchhalter, Vasilisa Rudneva, Ji Wan Park, Eun Pyo Hong, Seong Gu Heo, Anthony DiBiase, Kuan-lin Huang, Ivica Letunic, Michael D. McLellan, Steven J. Newhouse, Matthias Schlesner, Tal Shmaya, Sushant Kumar, David C. Wedge, Mark H. Wright, Venkata D. Yellapantula, Mark Gerstein, Ekta Khurana, Tomas Marques-Bonet, Arcadi Navarro, Carlos D. Bustamante, Jared T. Simpson, Li Ding, Reiner Siebert, Hidewaki Nakagawa, Douglas F. Easton; led by Stephan Ossowski, Jose M. C. Tubio, Gad Getz, Francisco M. De La Vega, Xavier Estivill, Jan O. Korbel. RNA-sequencing analysis: major contributions from Nuno A. Fonseca, Andre Kahles, Kjong-Van Lehmann, Lara Urban, Cameron M. Soulette, Yuichi Shiraishi, Fenglin Lu, Yao He, Deniz Demircioglu, Natalie R. Davidson, Claudia Calabrese, Junjun Zhang, Marc D. Perry, Qian Xiang; further contributions from Liliana Greger, Siliang Li, Dongbing Liu, Stefan G. Stark, Fan Zhang, Samirkumar B. Amin, Peter Bailey, Aurelien Chateigner, Isidro Cortes-Ciriano, Brian Craft, Serap Erkek, Milana Frenkel-Morgenstern, Mary Goldman, Katherine A. Hoadley, Yong Hou, Matthew R. Huska, Ekta Khurana, Helena Kilpinen, Jan O. Korbel, Fabien C. Lamaze, Chang Li, Xiaobo Li, Xinyue Li, Xingmin Liu, Maximilian G. Marin, Julia Markowski, Tannistha Nandi, Morten Muhlig Nielsen, Akinyemi I. Ojesina, Qiang Pan-Hammarstrom, Peter J. Park, Chandra Sekhar Pedamallu, Jakob Skou Pedersen, Reiner Siebert, Hong Su, Patrick Tan, Bin Tean Teh, Jian Wang, Sebastian M. Waszak, Heng Xiong, Sergei Yakneen, Chen Ye, Christina Yung, Xiuqing Zhang, Liangtao Zheng, Jingchun Zhu, Shida Zhu, Philip Awadalla, Chad J. Creighton, Matthew Meyerson, B. F. Francis Ouellette, Kui Wu, Huanming Yang; led by Jonathan Goke, Roland F. Schwarz, Oliver Stegle, Zemin Zhang, Alvis Brazma, Gunnar Ratsch, Angela N. Brooks. Clustering of tumour genomes based on telomere maintenance-related features: major contribution from David Haan; led by Lincoln D. Stein, Joshua M. Stuart. Clustered mutational processes in PCAWG: major contributions from Jonas Demeulemeester, Maxime Tarabichi, Matthew W. Fittall; led by Peter J. Campbell, Jan O. Korbel, Peter Van Loo. Tumours without detected driver mutations: Esther Rheinbay, Amaro Taylor-Weiner, Radhakrishnan Sabarinathan, Peter J. Campbell, Gad Getz. Panorama of driver mutations in human cancer: major contributions from Radhakrishnan Sabarinathan, Oriol Pich; further contributions from Inigo Martincorena, Carlota Rubio-Perez, Malene Juul, Jeremiah A. Wala, Steven Schumacher, Ofer Shapira, Nikos Sidiropoulos, Sebastian M. Waszak, David Tamborero, Loris Mularoni, Esther Rheinbay, Henrik Hornshøj, Jordi Deu-Pons, Ferran Muiños, Johanna Bertl, Qianyun Guo, Chad J. Creighton, Joachim Weischenfeldt, Jan O. Korbel, Gad Getz, Peter J. Campbell, Jakob Skou Pedersen, Rameen Beroukhim; led by Abel Gonzalez-Perez. Pilot benchmarking, variant consensus development and validation: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Yu Fan, David Torrents, Matthias Bieg, Paul C. Boutros, Ken Chen, Zechen Chong, Kristian Cibulskis, Oliver Drechsel, Roland Eils, Robert S. Fulton, Josep

# Article

Gelpi, Mark Gerstein, Santiago Gonzalez, Gad Getz, Ivo G. Gut, Faraz Hach, Michael Heinold, Taobo Hu, Vincent Huang, Barbara Hutter, Hyung-Lae Kim, Natalie Jager, Jongsun Jung, Sushant Kumar, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Ivica Letunic, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, R. Jay Mashl, Michael D. McLellan, Ana Milovanovic, Morten Muhlig Nielsen, Brian D. O'Connor, Stephan Ossowski, Nagarajan Paramasivam, Jakob Skou Pedersen, Marc D. Perry, Montserrat Puiggros, Romina Royo, Esther Rheinbay, S. Cenik Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Grace Tiao, Jeremiah A. Wala, Jiayin Wang, Wenyi Wang, Sebastian M. Waszak, Joachim Weischenfeldt, Michael Wendl, Johannes Werner, Zhenggang Wu, Hong Xue, Sergei Yakneen, Takafumi N. Yamaguchi, Kai Ye, Venkata Yellapantula, Junjun Zhang, David A. Wheeler; led by Li Ding, Jared T. Simpson. Production somatic variant calling on the PCAWG compute cloud: major contributions from Christina K. Yung, Brian D. O'Connor, Sergei Yakneen, Junjun Zhang; further contributions from Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M. Raine, Romina Royo, Gordon Saksena, Matthias Schlesner, Solomon I. Shorser, Miguel Vazquez, Joachim Weischenfeldt, Denis Yuen, Adam P. Butler, Brandi N. Davis-Dusenbery, Roland Eils, Vincent Ferretti, Robert L. Grossman, Olivier Harismendy, Youngwook Kim, Hidewaki Nakagawa, Steven J Newhouse, David Torrents; led by Lincoln D. Stein. PCAWG data portals: major contributions from Mary Goldman, Junjun Zhang, Nuno A. Fonseca, Isidro Cortes-Ciriano; further contributions from Qian Xiang, Brian Craft, Elena Pineiro-Yanez, Brian D O'Connor, Wojciech Bazant, Elisabet Barrera, Alfonso Munoz, Robert Petryszak, Anja Fullgrabe, Fatima Al-Shahrour, Maria Keays, David Haussler, John Weinstein, Wolfgang Huber, Alfonso Valencia, Irene Papatheodorou, Jingchun Zhu; led by Vincent Ferretti, Miguel Vazquez.

**Competing interests** Gad Getz receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMutSig and POLYSOLVER. Hikmat Al-Ahmadie is consultant for AstraZeneca and Bristol-Myers Squibb. Samuel Aparicio is a founder and shareholder of Contextual Genomics. Pratiti Bandopadhyay receives grant funding from Novartis for an unrelated project. Rameen Beroukhi owns equity in Ampressa Therapeutics. Andrew Biankin receives grant funding from Celgene, AstraZeneca and is a consultant for or on advisory boards of AstraZeneca, Celgene, Elstar Therapeutics, Clovis Oncology and Roche. Ewan Birney is a consultant for Oxford Nanopore, Dovetail and GSK. Marcus Bosenberg is a consultant for Eli Lilly. Atul Butte is a cofounder of and consultant for Personalis, NuMedii, a consultant for Samsung, Geisinger Health, Mango Tree Corporation, Regentrief Institute and in the recent past a consultant for 10x Genomics and Helix, a shareholder in Personalis, a minor shareholder in Apple, Twitter, Facebook, Google, Microsoft, Sarepta, 10x Genomics, Amazon, Biogen, CVS, Illumina, Snap and Sutro and has received honoraria and travel reimbursement for invited talks from Genentech, Roche, Pfizer, Optum, AbbVie and many academic institutions and health systems. Carlos Caldas has served on the Scientific Advisory Board of Illumina. Lorraine Chantrill acted on an advisory board for AMGEN Australia in the past 2 years. Andrew D. Cherniack receives research funding from Bayer. Helen Davies is an inventor on a number of patent applications that encompass the use of mutational signatures. Francisco De La Vega was employed at Annai Systems during part of the project. Ronny Drapkin serves on the scientific advisory board of Repare Therapeutics and Siamab Therapeutics. Rosalind Eeles has received an honorarium for the GU-ASCO meeting in San Francisco in January 2016 as a speaker, a honorarium and support from Janssen for the RMH FR meeting in November 2017 as a speaker (title: genetics and prostate cancer), a honorarium for an University of Chicago invited talk in May 2018 as speaker and an educational honorarium paid by Bayer & Ipsen to attend GU Connect 'Treatment sequencing for mCRPC patients within the changing landscape of mHSPC' at a venue at ESMO, Barcelona, on 28 September 2019. Paul Flicek is a member of the scientific advisory boards of Fabric Genomics and Eagle Genomics. Ronald Ghossein is a consultant for Veracyte. Dominik Glodzik is an inventor on a

number of patent applications that encompass the use of mutational signatures. Eoghan Harrington is a full-time employee of Oxford Nanopore Technologies and is a stock holder. Yann Joly is responsible for the Data Access Compliance Office (DACO) of ICGG 2009-2018. Sissel Juul is a full-time employee of Oxford Nanopore Technologies and is a stock holder. Vincent Khoo has received personal fees and non-financial support from Accuray, Astellas, Bayer, Boston Scientific and Janssen. Stian Knappskog is a coprincipal investigator on a clinical trial that receives research funding from AstraZeneca and Pfizer. Ignaty Leshchiner is a consultant for PACT Pharma. Carlos López-Otín has ownership interest (including stock and patents) in DREAMgenics. Matthew Meyerson is a scientific advisory board chair of, and consultant for, Origimed, has obtained research funding from Bayer and Ono Pharma and receives patent royalties from LabCorp. Serena Nik-Zainal is an inventor on a number of patent applications that encompass the use of mutational signatures. Nathan Pennell has done consulting work with Merck, Astrazeneca, Eli Lilly and Bristol-Myers Squibb. Xose S. Puente has ownership interest (including stock and patents in DREAMgenics. Benjamin J. Raphael is a consultant for and has ownership interest (including stock and patents) in Medley Genomics. Jorge Reis-Filho is a consultant for Goldman Sachs and REPARE Therapeutics, member of the scientific advisory board of Volition RX and Paige.AI and an ad hoc member of the scientific advisory board of Ventana Medical Systems, Roche Tissue Diagnostics, Invivo, Roche, Genentech and Novartis. Lewis R. Roberts has received grant support from ARIAD Pharmaceuticals, Bayer, BTG International, Exact Sciences, Gilead Sciences, Glycotest, RedHill Biopharma, Target PharmaSolutions and Wako Diagnostics and has provided advisory services to Bayer, Exact Sciences, Gilead Sciences, GRAIL, QED Therapeutics and TAVEC Pharmaceuticals. Richard A. Scolyer has received fees for professional services from Merck Sharp & Dohme, GlaxoSmithKline Australia, Bristol-Myers Squibb, Dermepedia, Novartis Pharmaceuticals Australia, Myriad, NeraCare GmbH and Amgen. Tal Shmaya is employed at Annai Systems. Reiner Siebert has received speaker honoraria from Roche and AstraZeneca. Sabina Signoretti is a consultant for Bristol-Myers Squibb, AstraZeneca, Merck, AACR and NCI and has received funding from Bristol-Myers Squibb, AstraZeneca, Exelixis and royalties from Biogenex. Jared Simpson has received research funding and travel support from Oxford Nanopore Technologies. Anil K. Sood is a consultant for Merck and Kiyatec, has received research funding from M-Trap and is a shareholder in BioPath. Simon Tavaré is on the scientific advisory board of Ipsen and a consultant for Kallyope. John F. Thompson has received honoraria and travel support for attending advisory board meetings of GlaxoSmithKline and Provectus and has received honoraria for participation in advisory boards for MSD Australia and BMS Australia. Daniel Turner is a full-time employee of Oxford Nanopore Technologies and is a stock holder. Naveen Vasudev has received speaker honoraria and/or consultancy fees from Bristol-Myers Squibb, Pfizer, EUSA pharma, MSD and Novartis. Jeremiah A. Wala is a consultant for Nference. Daniel J. Weisenberger is a consultant for Zymo Research. Dai-Ying Wu is employed at Annai Systems. Cheng-Zhong Zhang is a cofounder and equity holder of Pillar Biosciences, a for-profit company that specializes in the development of targeted sequencing assays. The other authors declare no competing interests.

## Additional information

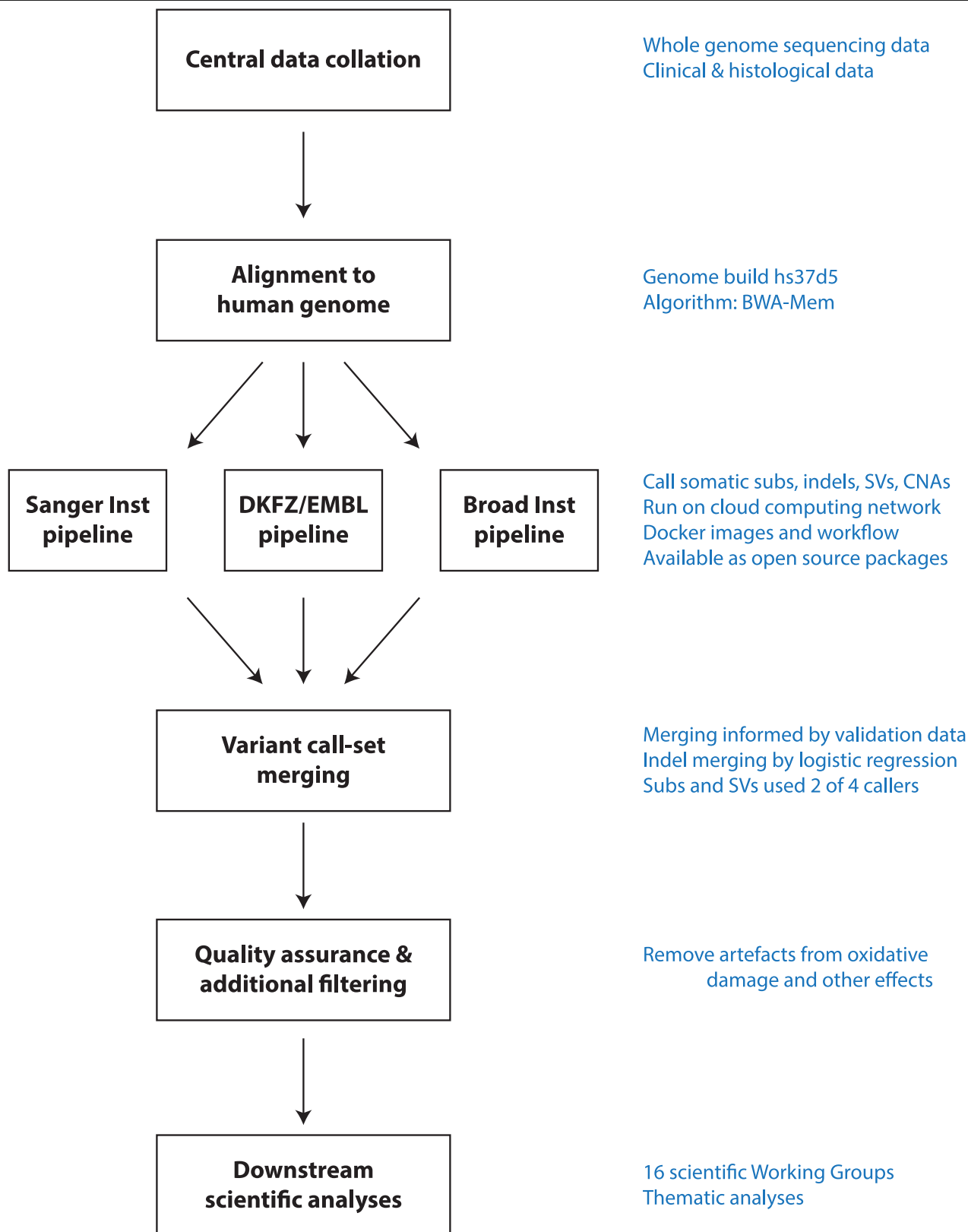
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1969-6>.

**Correspondence and requests for materials** should be addressed to P.J.C., G.G., J.O.K., J.M.S. or L.D.S.

**Peer review information** *Nature* thanks Arul Chinnaiyan, Ben Lehner, Nicolas Robine and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

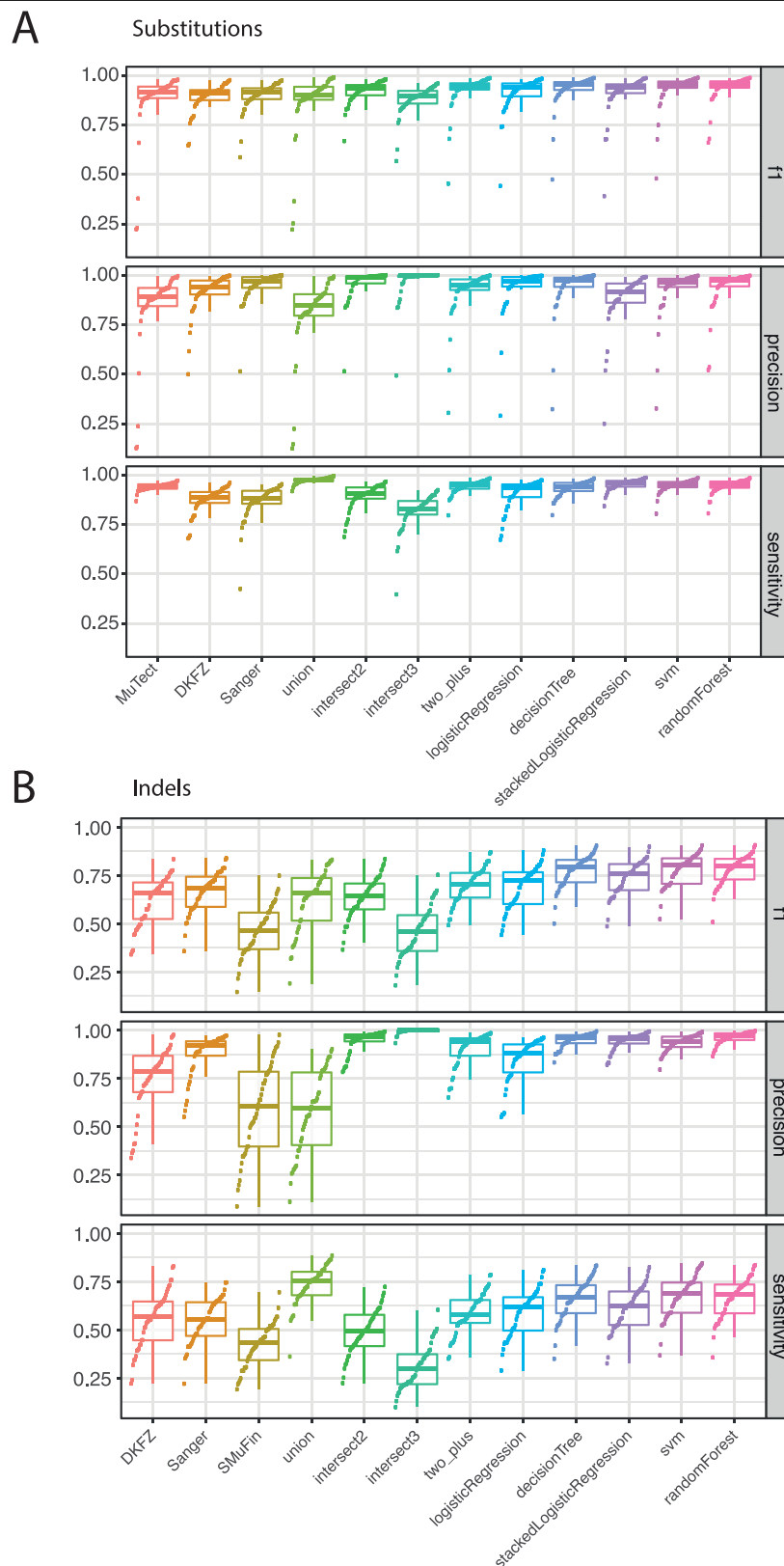
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





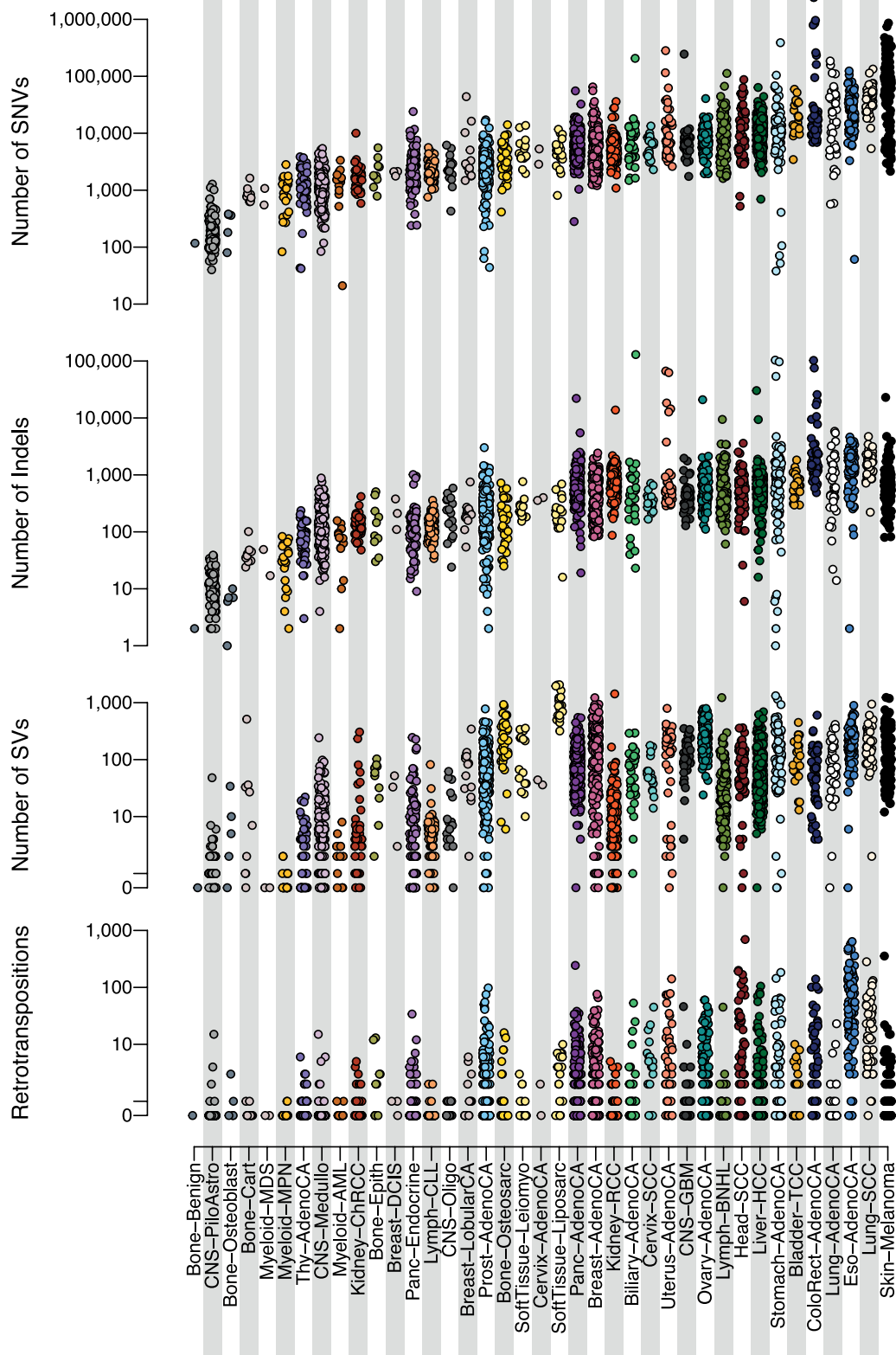
**Extended Data Fig. 1 | Flow-chart showing key steps in the analysis of PCAWG genomes.** After alignment to the genome, somatic mutations were identified by three pipelines, with subsequent merging into a consensus variant set used

for downstream scientific analyses. Subs, substitutions; DKFZ/EMBL, the German Cancer Research Centre (DKFZ) and European Molecular Biology Laboratory (EMBL).



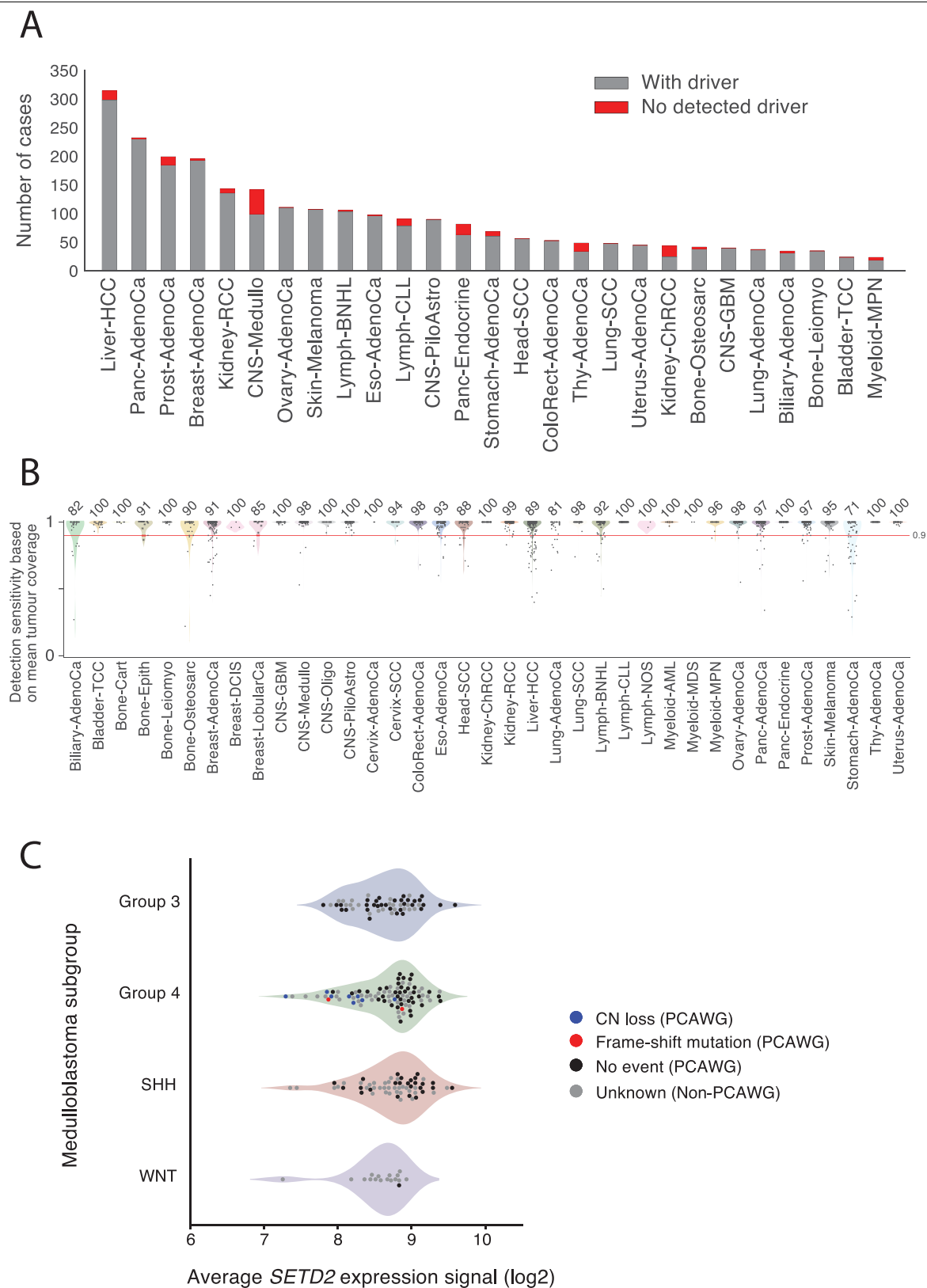
**Extended Data Fig. 2 | Distribution of accuracy estimates across algorithms and samples from validation data. a.**  $F_1$  accuracy, precision and sensitivity estimates for somatic SNVs across the core algorithms and different approaches to merging the call-sets. The box plots demarcate the interquartile range and median of estimates across the  $n = 50$  samples in the validation

dataset. **b.**  $F_1$  accuracy, precision and sensitivity estimates for somatic indels ( $n = 50$  samples). SVM, support vector machine; union, calls made by all variant-calling algorithms; intersect2, calls made by any combination of two variant-calling algorithms; intersect3, calls made by any three variant-calling algorithms.



**Extended Data Fig. 3 | Distribution of numbers of somatic mutations of different classes across tumour types.** The y axis is on a log scale. The 2,583 donors with the highest quality metrics (white-listed donors) are plotted. SNVs

indicate substitutions; indels are taken as insertions or deletions <100 bp in size; retrotranspositions are the combined counts of somatic retrotransposon insertions, transductions and somatic pseudogene insertions.

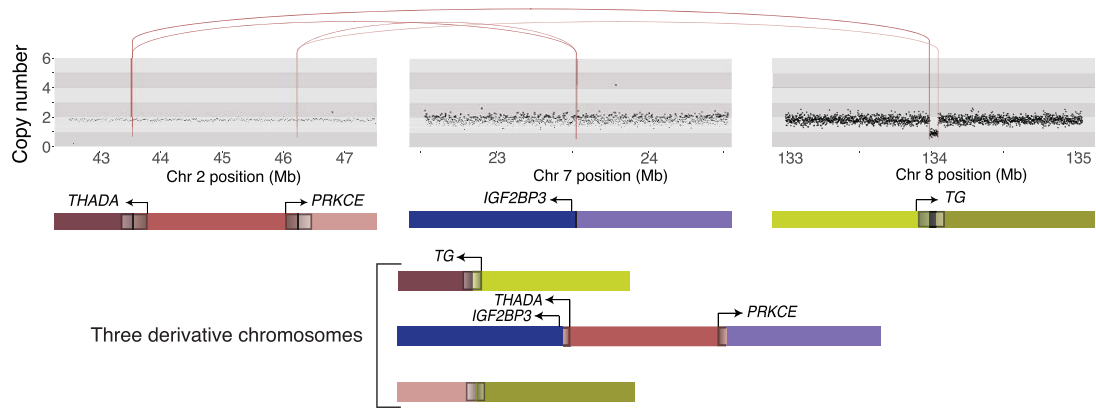


**Extended Data Fig. 4 | Patients with no detected driver mutations in PCAWG.**  
**a**, Number (red) of patients without detected driver mutations distributed across the different tumour types studied. **b**, Estimated sensitivity for detecting somatic point mutations genome-wide across tumour types (total sample size:  $n = 2,583$  patients). Each point represents the estimate for a single patient, layered on violin plots that show the estimated density distribution of

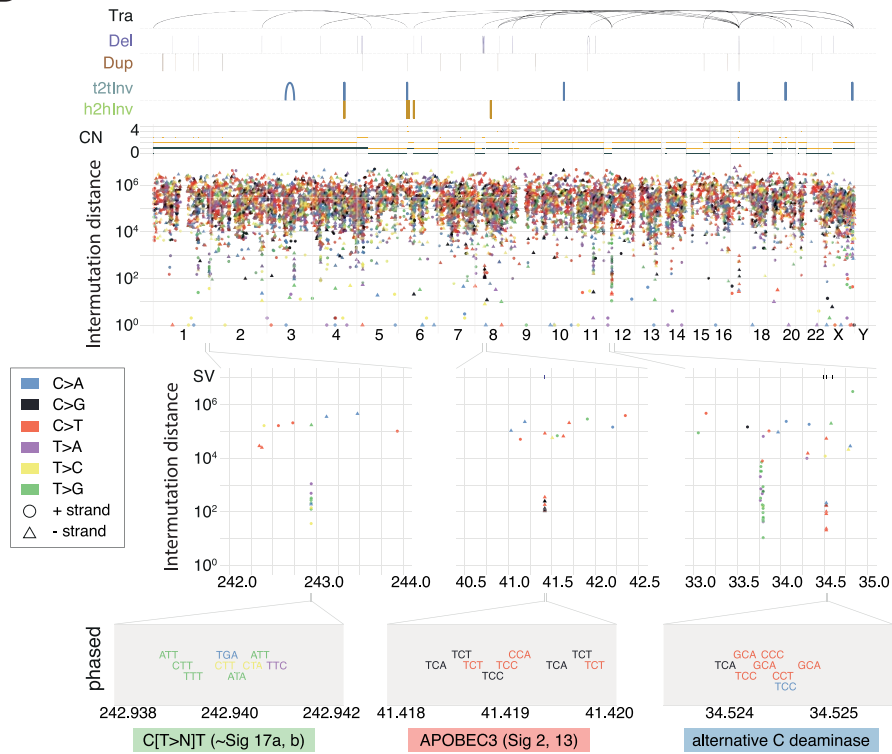
sensitivity values for that tumour type (the width proportional is to density). **c**, *SETD2* expression levels across different medulloblastoma subtypes. Points represent individual patients, coloured by whether the gene exhibited focal copy number (CN) loss or a truncating point mutation, or was the wild-type gene. The coloured areas are violin plots showing the estimated density distribution of expression values for that medulloblastoma subtype.



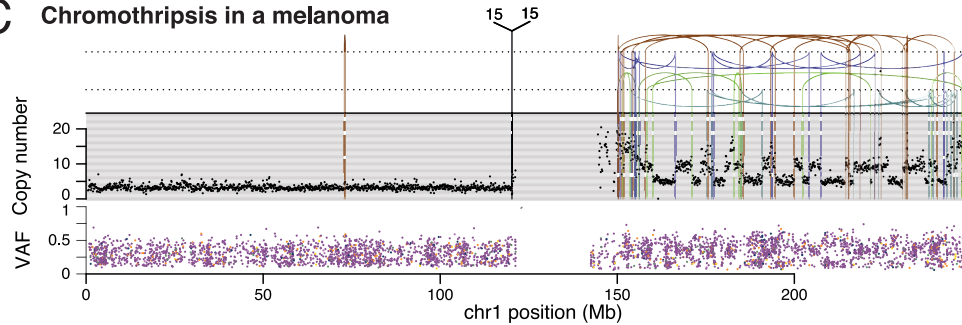
## A Chromoplexy in a thyroid adenocarcinoma



## B Kataegis in a pancreatic adenocarcinoma



## C Chromothripsis in a melanoma

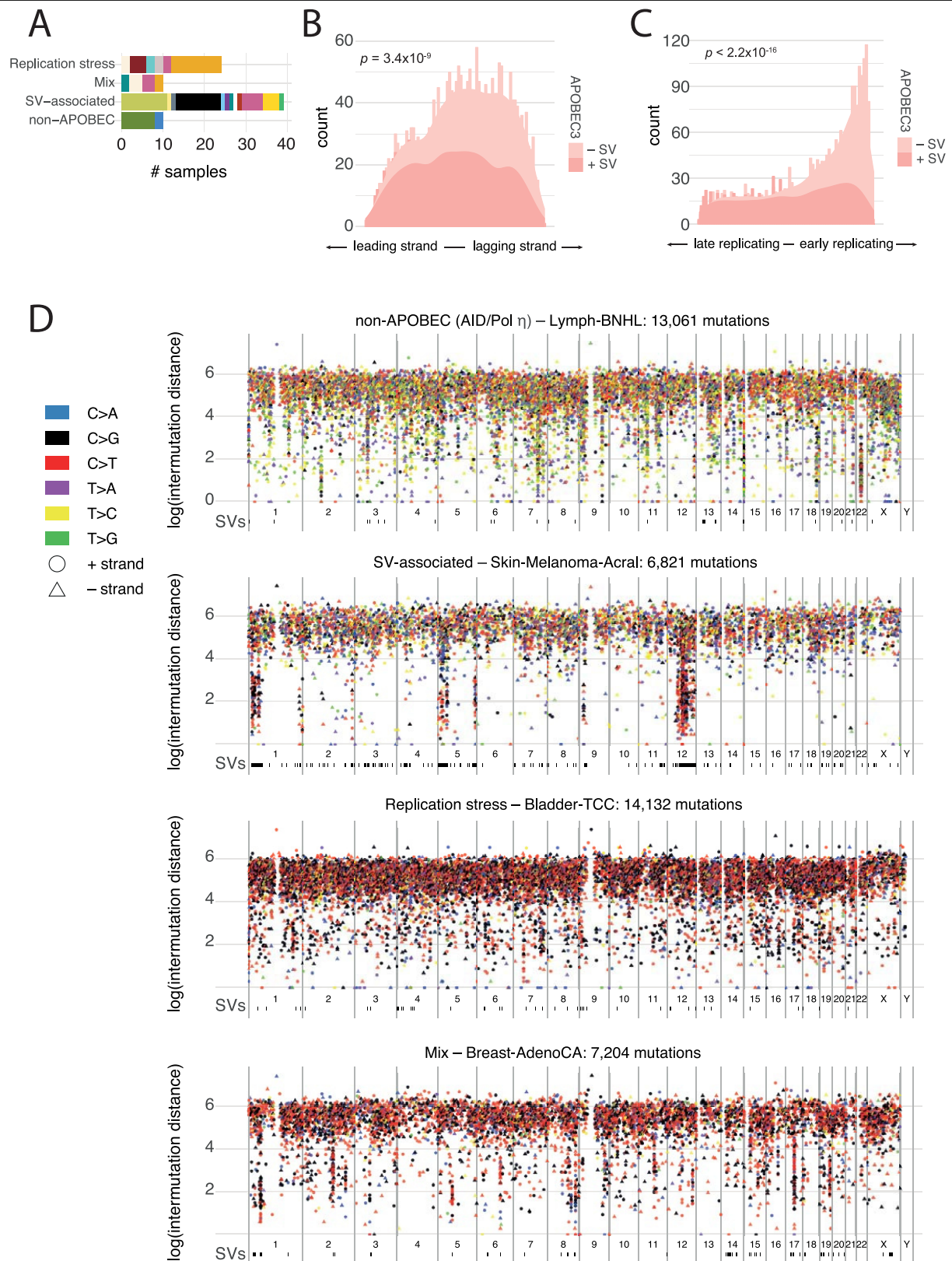


Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | Examples of clustered mutational processes.**

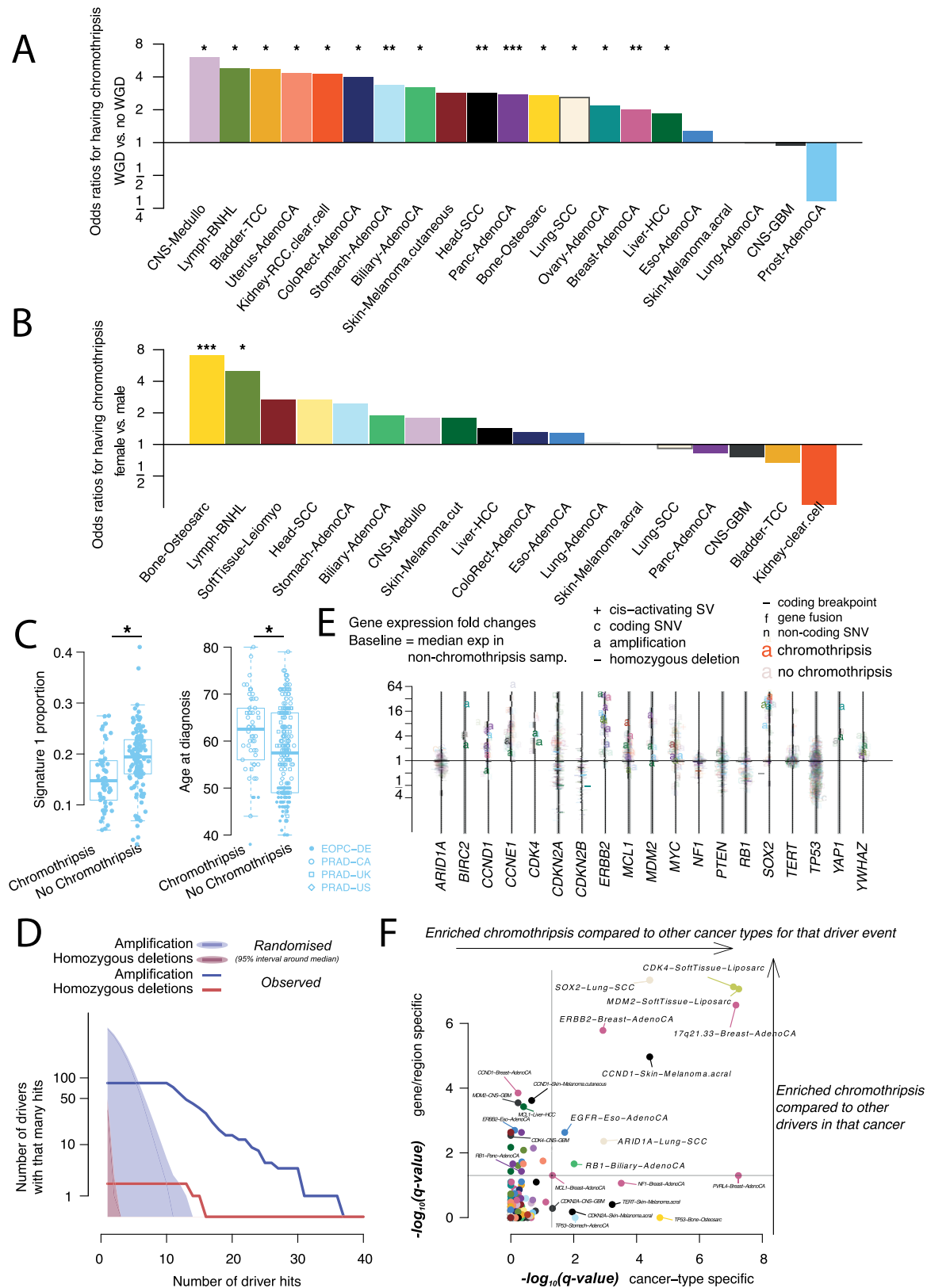
**a**, Chromoplexy example in a thyroid adenocarcinoma. Genes at the breakpoints are schematically depicted in their normal genomic context and again in the reconstructed derivative chromosomes below. **b**, Distinct kataegis signatures in the genome of a pancreatic adenocarcinoma sample. SVs and their classification are shown above the main rainfall plot, as well as the total and minor allele copy number. Tra, translocation; del, deletion; dup, duplication; t2tInv, tail-to-tail inversion; h2hInv, head-to-head inversion. Magnifications of the three foci on chromosomes 1, 8 and 12, respectively, highlight distinct manifestations of kataegis. Left, a novel process similar to signature 17 with T > N mutations at CT or TT dinucleotides. Middle, the

prototypical APOBEC3A/B type with C > T (signature 2) and/or C > G/A (signature 13) substitutions at TpC. Right, an alternative cytidine deaminase(s) with a preference for substitutions at C/GpC. Most of the SNVs in each of these foci can be phased to the same allele and no evidence of anti-phasing is observed. **c**, Example of a chromothripsis event in a melanoma. The black points (top) represent copy-number estimates from individual genomic bins, with SVs shown as coloured arcs (translocation in black, deletion in purple, duplication in brown, tail-to-tail inversion in cyan, head-to-head inversion in green) that mostly demarcate copy-number changes. The mate chromosomes are displayed above translocations. Bottom, the variant allele fractions of somatic mutations distributed along the relevant chromosomal region.



**Extended Data Fig. 6 | Patterns of intense kataegis.** **a**, Distribution of the tumour types (colour-coded as in Extended Data Fig. 3) of the samples in the top 5% of kataegis intensity in each of the four identified genome-wide patterns: non-APOBEC, replication stress, rearrangement-associated and the combination of the last two. **b**, **c**, Distribution of leading/lagging strand (**b**) and

replication timing bias (**c**) for rearrangement-(in)dependent APOBEC kataegis, based on  $n = 2,583$  tumours.  $P$  values were derived using a two-sided Mann-Whitney  $U$ -test. **d**, Example rainfall plots for each of the four identified kataegis patterns.

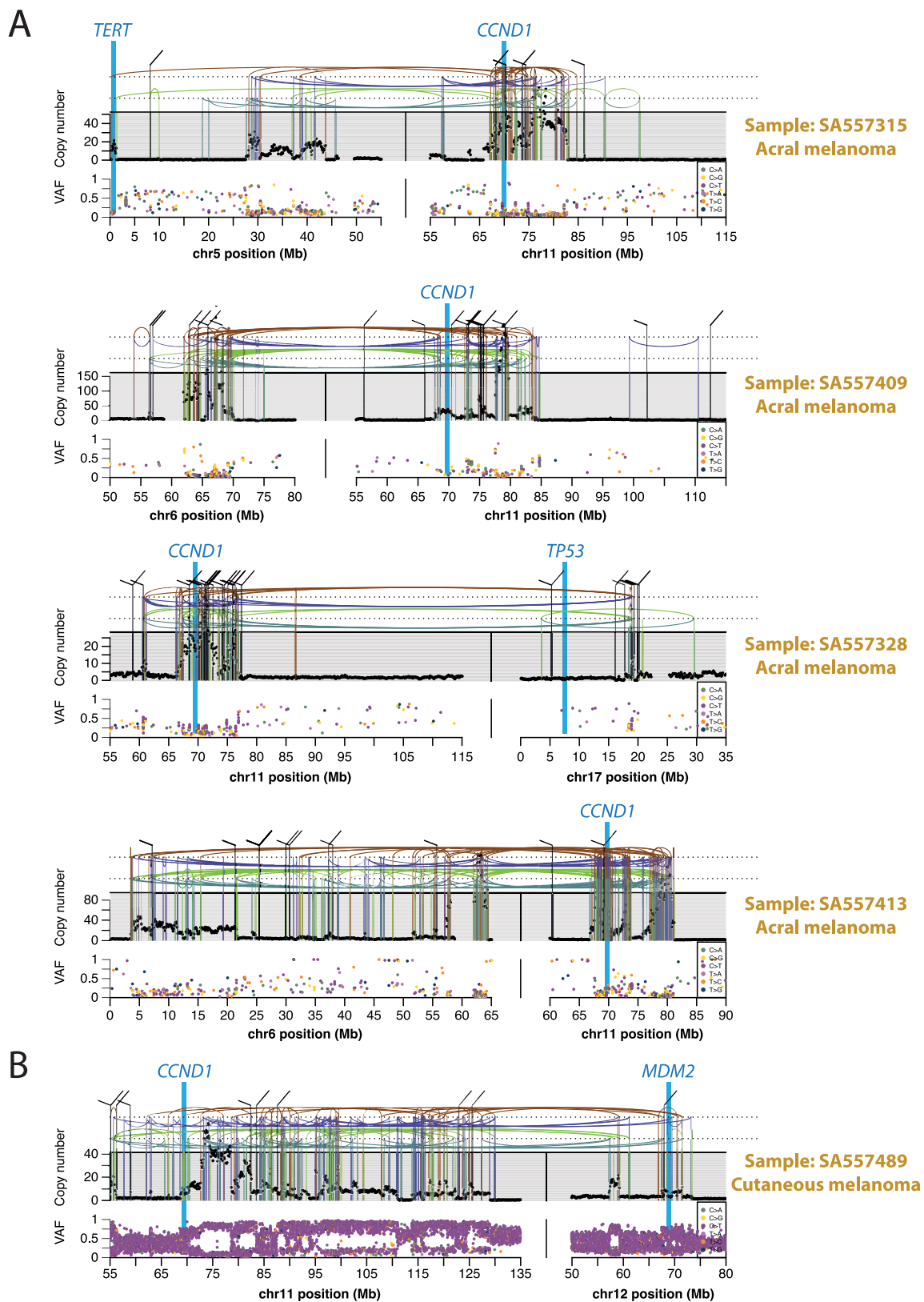


Extended Data Fig. 7 | See next page for caption.



**Extended Data Fig. 7 | Association of chromothripsis with covariates and driver events.** **a**, Odds ratios per cancer type of containing chromothripsis in whole-genome duplicated versus diploid samples ( $n = 2,583$  patients). \*\*\* $q < 0.001$ ; \*\* $q < 0.01$ ; \* $q < 0.05$ . Two-sided hypothesis testing was performed using Fisher–Boschloo tests, corrected for multiple-hypothesis testing. **b**, Same as **a** for female versus male. **c**, Proportion of mutations explained by single-base substitution signature 1 and age at diagnosis in prostate cancer samples ( $n = 210$  patients) with or without chromothripsis ( $q < 0.05$ ). The early-onset prostate cancer project drives the signal and was sequenced at lower depth. For the box-and-whisker plots, the box denotes the interquartile range, with the median marked as a horizontal line. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Two-sided hypothesis testing was performed using Mann–Whitney  $U$ -tests. **d**, Counts of co-occurrence of chromothripsis with amplification (blue) and homozygous deletions (red) in driver regions: observed (thick line) versus randomized (shaded area and thin line). The cumulative number of drivers that were hit is

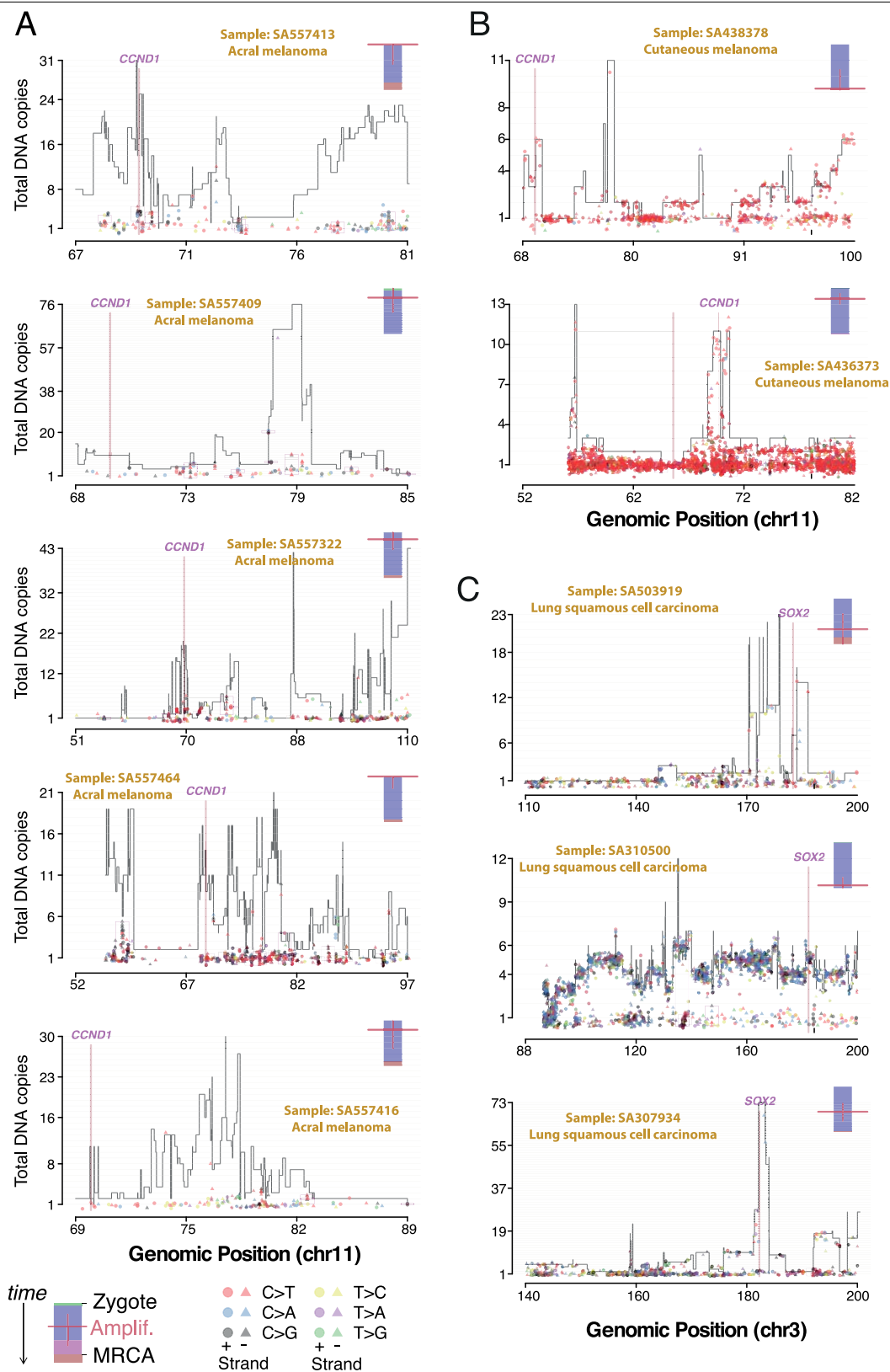
plotted as a function of the number of times those drivers were hit. **e**, For each sample in which chromothripsis coincided with a driver event in those genes, we show the fold change in gene expression compared to the median expression of the gene in non-chromothripsis samples of the same cancer type, coloured by cancer type and shaped by the type of driver event. We show with added transparency the fold changes calculated the same way for samples with driver mutations hitting the same driver genes, but that had no evidence of chromothripsis. Analysis is based on  $n = 1,222$  patients with RNA-sequencing data. **f**, Enrichment of co-occurrence of chromothripsis with driver events. The  $x$  axis shows the association of chromothripsis with a driver in a given cancer type compared with its rate of association with that driver in all other cancer types. The  $y$  axis shows the association of chromothripsis with a driver in a given cancer type compared with its rate of association with all other drivers in that type. Exact binomial tests are used and  $P$  values are corrected for multiple testing according to the Benjamini–Hochberg method.



Extended Data Fig. 8 | See next page for caption.

**Extended Data Fig. 8 | Further examples of chromothripsis-induced amplification targeting multiple cancer-associated genes simultaneously in melanoma.** **a**, Examples of amplifications that occurred early in the development of melanoma. The black points (top) represent copy-number estimates from individual genomic bins, with SVs shown as coloured arcs (translocation in black, deletion in purple, duplication in brown, tail-to-tail inversion in cyan and head-to-head inversion in green) that mostly demarcate copy-number changes. Bottom, the variant allele fractions of SNVs distributed

along the relevant chromosomal region. The paucity of somatic mutations at high variant allele fractions in the most-heavily amplified regions indicates that these amplifications began very early in tumour evolution, before the lineage had had opportunity to acquire many SNVs. **b**, Example of an amplification that occurred late in melanoma development. The large numbers of somatic mutations at high variant allele fractions in the most-heavily amplified regions indicate that these amplifications began late in tumour evolution, after the lineage had already acquired many SNVs.

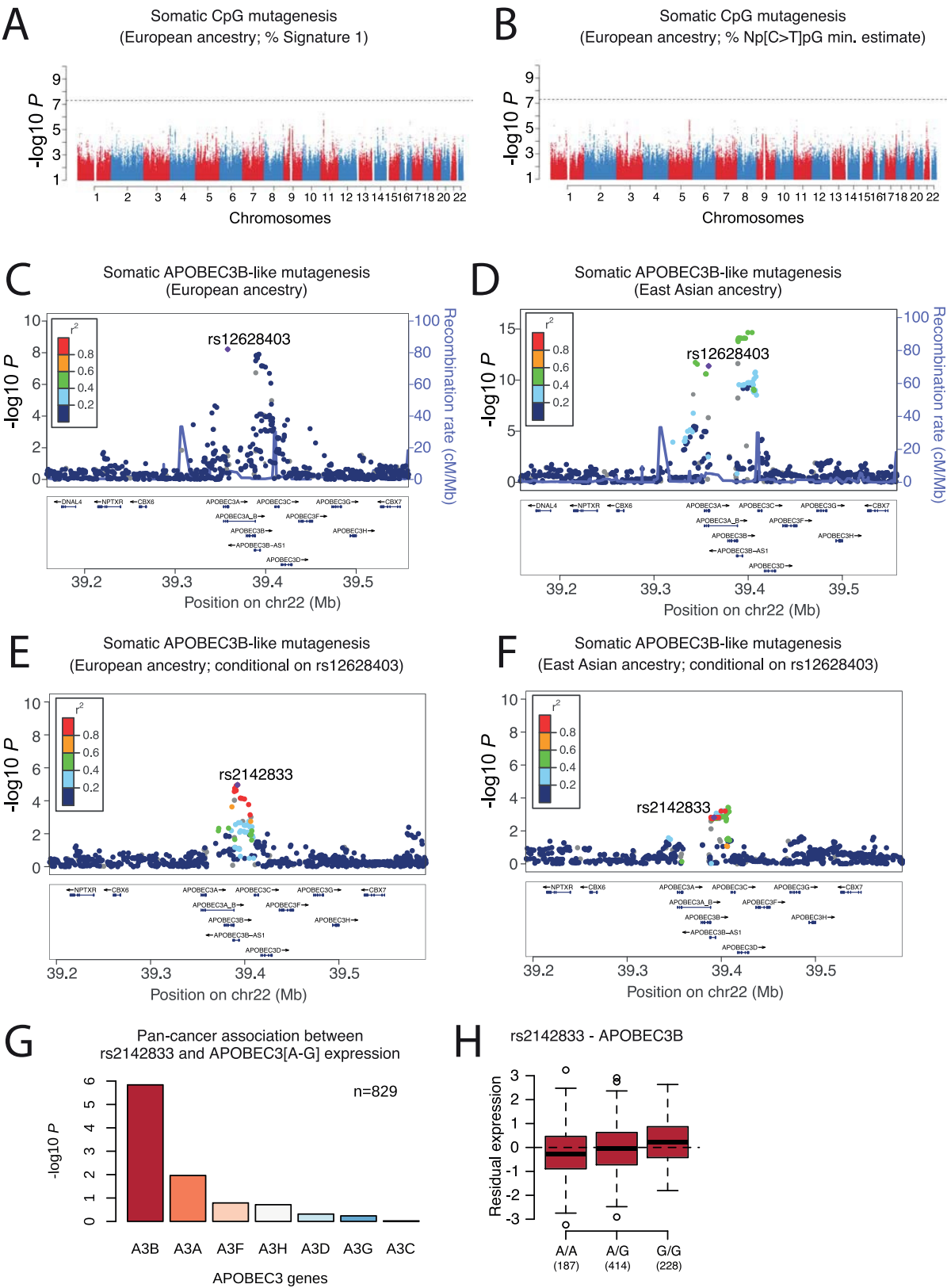


Extended Data Fig. 9 | See next page for caption.



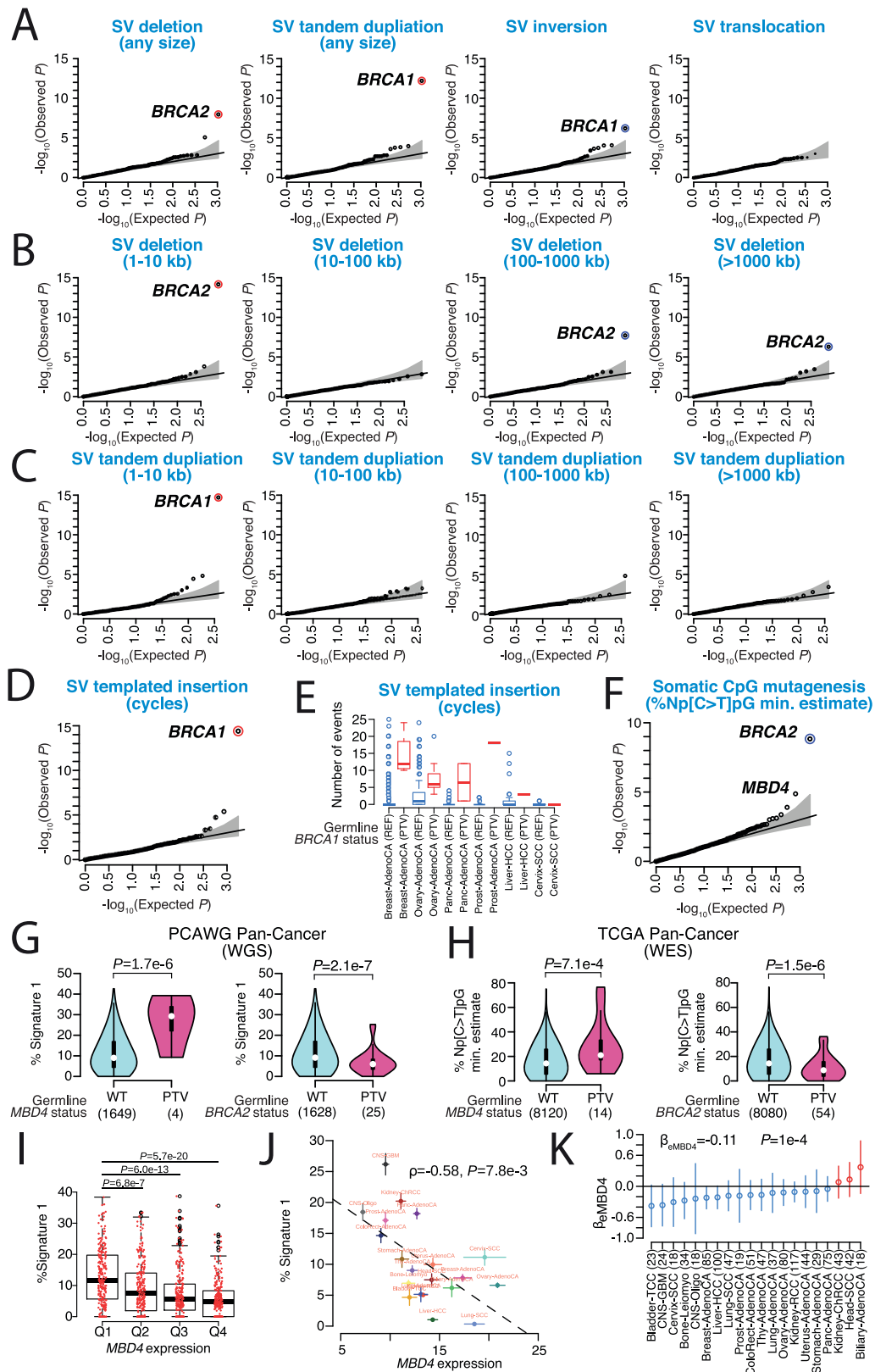
**Extended Data Fig. 9 | Timing the amplifications after chromothripsis in molecular time for 10 representative cases. a,** Copy-number plot of chromothriptic regions categorized as 'liposarc-like' in five acral melanomas with *CCND1* amplification. Segments indicate the copy number of the major allele. Points represent SNV multiplicities, that is, the estimated number of copies carrying each SNV, coloured by base change and shaped by strand. Small vertical arrows link SNVs to their corresponding copy-number segment. Kataegis foci are shown within black boxes and show typical strand specificities (all triangles or all circles), similar multiplicities and base changes of signatures 2 and 13 (red and black, respectively). A coloured bar (top right)

represents the molecular timing of the amplification (red bar; high is early, low is late) and is coloured by the fraction of total SNVs assigned to the following timing categories: clonal [early], clonal mutations that occurred before duplications involving the relevant chromosome (including whole-genome duplications); clonal [late], clonal mutations that occurred after such duplications; and clonal [NA], mutations that occurred when no duplication was observed. **b,** Same as **a** in two cutaneous melanomas, one shows early amplification, the other late amplification. **c,** Same as **a, b**, for three lung squamous cell carcinomas and late amplification of *SOX2*.



**Extended Data Fig. 10 | Association between common germline variants and endogenous mutational processes.** Genome-wide association of somatic CpG mutagenesis in individuals of European ancestry ( $n = 1,201$  patients) based on mutational signature analysis (a) and NpCpG motif analysis (b). Two-sided hypothesis testing was performed using PLINK v.1.9. To mitigate multiple-hypothesis testing, the significance threshold was set to genome-wide significance ( $P < 5 \times 10^{-8}$ ). c, d, Locuszoom plot for somatic APOBEC3B-like mutagenesis association results, linkage disequilibrium and recombination rates around the genome-wide significant 22q13.1 locus in individuals with European (c) and East Asian (d) ancestry ( $n = 1,201$  and 318 patients,

respectively). Locuszoom plot for somatic APOBEC3B-like mutagenesis association results around the 22q13.1 locus in individuals with European (e) and East Asian (f) ancestry after conditioning on rs12628403. g, h, Association between rs2142833 and expression of *APOBEC3* genes in PCAWG tumour samples (adjusted for sex, age at diagnosis, histology and population structure in linear-regression models with two-sided hypothesis testing not corrected for multiple tests). For the box-and-whisker plot, the box denotes the interquartile range, with the median marked as a horizontal line. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Outliers are shown as points.

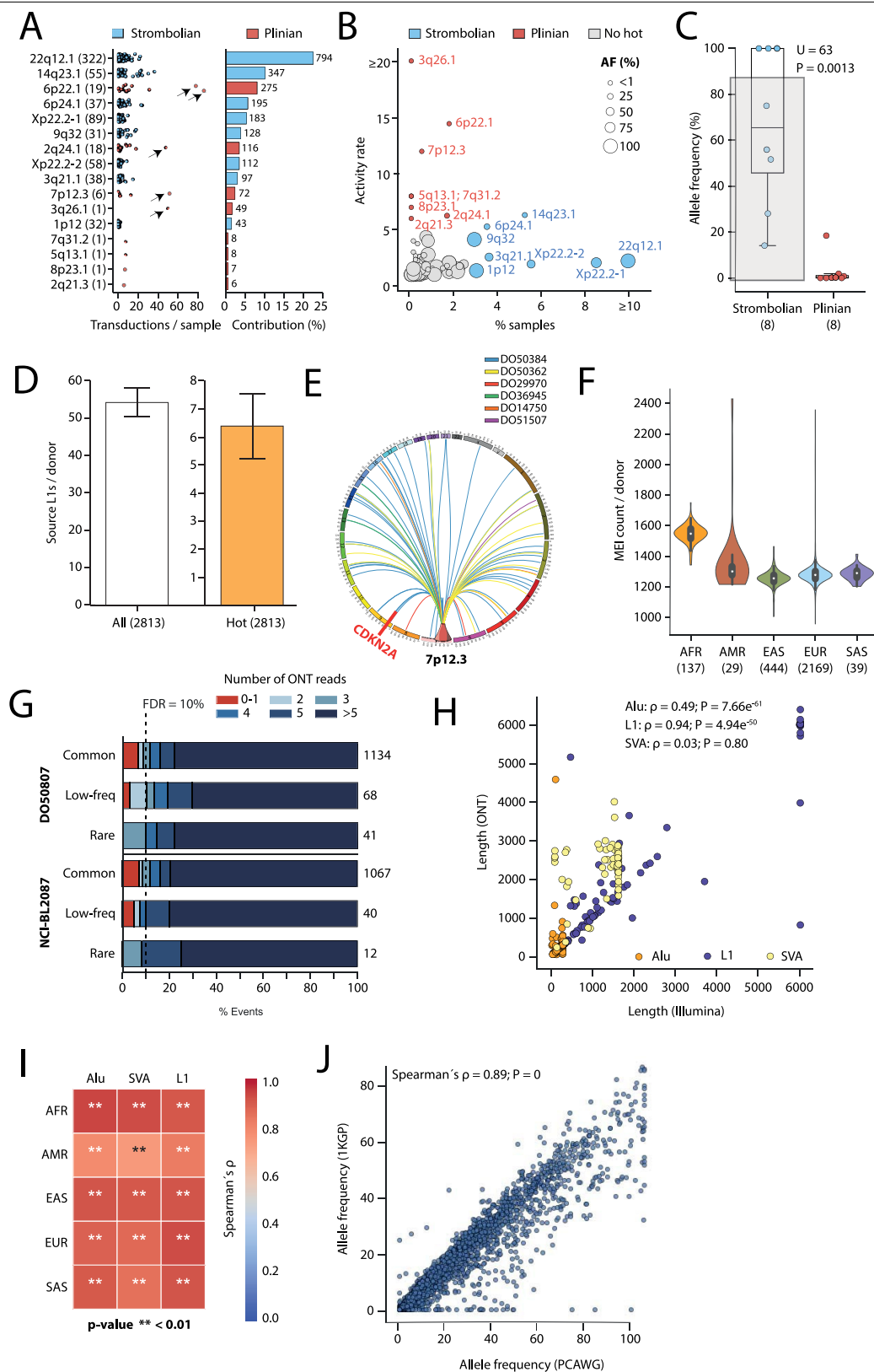


Extended Data Fig. 11 | See next page for caption.



**Extended Data Fig. 11 | Association between rare germline PTVs in protein-coding genes and somatic mutational phenotypes. a–d, f,** Data are based on two-sided rare-variant association testing across  $n = 2,583$  patients, with a stringent  $P$  value threshold of  $P < 2.5 \times 10^{-6}$  used to mitigate multiple-hypothesis testing (significant genes marked with coloured circles). Blue/red circles mark genes that decrease/increase somatic mutation rates. The black line represents the identity line that would be followed if the observed  $P$  values followed the null expectation, with the shaded area showing the 95% confidence intervals. **a,** QQ plots for the proportion of somatic SV deletions, tandem duplications, inversions and translocation in cancer genomes. **b,** QQ plots for the proportion of somatic SV deletions in cancer genomes stratified by four size groups (1–10 kb, 10–100 kb, 100–1,000 kb and >1,000 kb). **c,** QQ plots for the proportion of somatic SV tandem duplications in cancer genomes stratified by four size groups (1–10 kb, 10–100 kb, 100–1,000 kb and >1,000 kb). **d,** QQ plot for the presence or absence of somatic SV templated insertion (cycles) in cancer genomes. **e,** Number of SV-templated insertion cycles in PCAWG tumours with germline *BRCA1* PTVs. Only histological samples with at least one germline *BRCA1* PTV carrier are shown ( $n = 1,095$  patients combined). The box denotes the interquartile range, with the median marked as a horizontal line. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Outliers are shown as points. **f,** QQ plot for somatic CpG mutagenesis in cancer genomes based on NpCpG motif analysis. **g,** Violin plots show estimated densities of the proportion of somatic CpG mutations in PCAWG donors with germline *MBD4* and *BRCA2* PTVs. The box denotes the

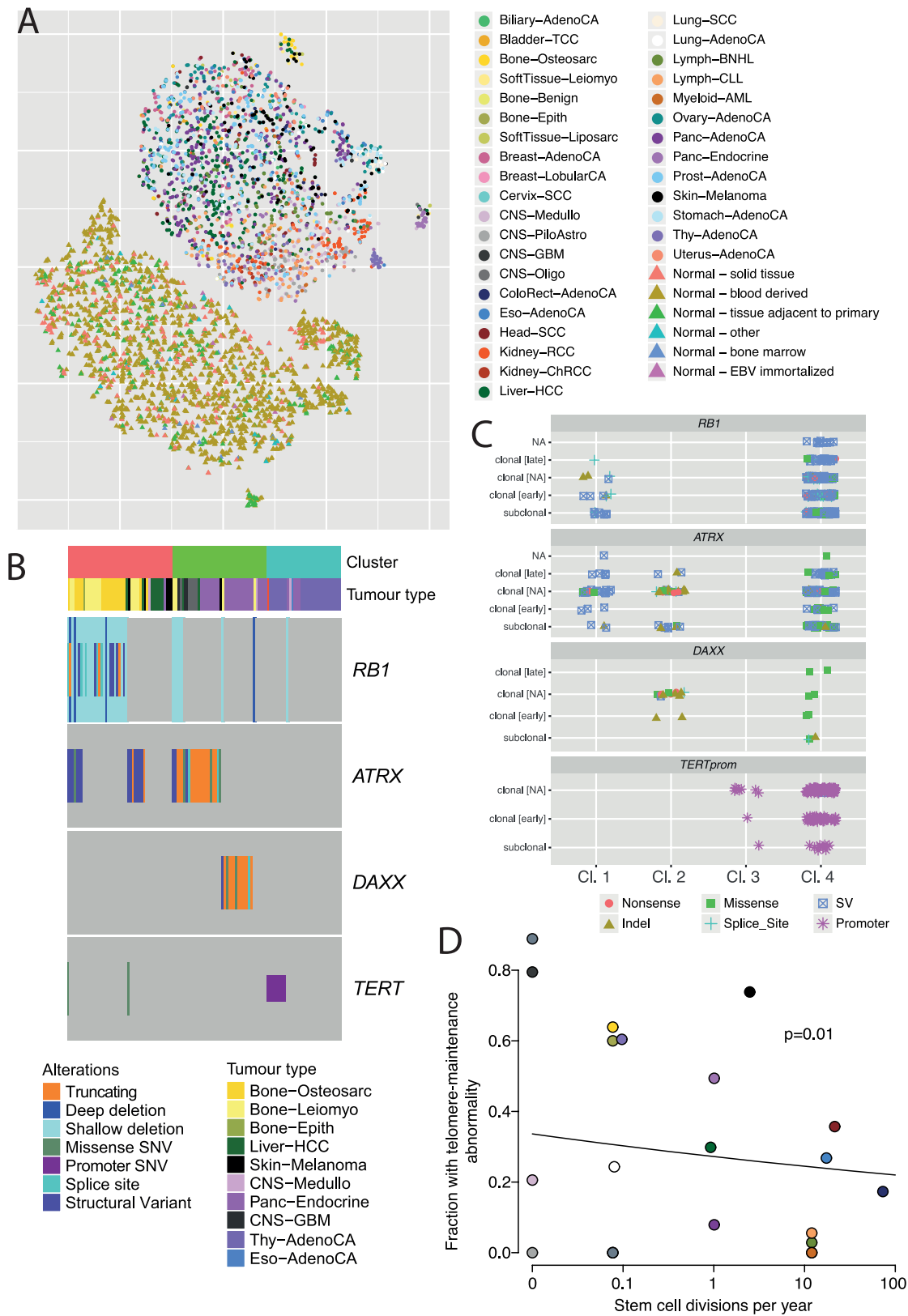
interquartile range, with the median marked as a white point. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Two-sided hypothesis testing, not corrected for multiple testing, was performed using linear regression models. **h,** Replication of germline *MBD4* and *BRCA2* PTV associations with somatic CpG mutagenesis in TCGA whole-exome sequencing donors. Violin plots show the estimated density of the proportion of somatic CpG mutations in TCGA exomes with germline *MBD4* and *BRCA2* PTVs. The box denotes the interquartile range, with the median marked as a white point. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Two-sided hypothesis testing, not corrected for multiple testing, was performed using linear-regression models. **i,** Correlation between *MBD4* expression and somatic CpG mutagenesis in primary solid PCAWG tumours. Hypothesis testing was two-sided and not corrected for multiple testing, using linear-regression models. The box denotes the interquartile range, with the median marked as a horizontal line. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. **j,** Data are mean  $\pm$  s.e.m. across  $n = 20$  tumour types. The dashed black line shows the fitted line to the data, estimated using linear-regression models. Hypothesis testing was two-sided and not corrected for multiple testing, using Spearman's rank correlations. **k,** *MBD4* effect sizes (open circles) with 95% confidence intervals (error bars) for individual cancer types were estimated using linear-regression analysis after (if available) accounting for sex, age at diagnosis (young/old) and ICGC project. Hypothesis testing was two-sided and not corrected for multiple testing.



Extended Data Fig. 12 | See next page for caption.

**Extended Data Fig. 12 | Germline MEI call set.** **a**, Left, dots show the number of transductions promoted by each hot element in individual samples. Arrows highlight retrotransposition burst. Right, the contribution of each hot locus is represented. The total number of transductions mediated by each source element is shown on the right. **b**, Source L1 activity rate (that is, measured as the average number of transductions mediated by an element) versus the percentage of samples with retrotransposition activity in which the germline element is active. For visualization purposes, extreme points observed for a source L1 with an activity rate of 49 and for a L1 active in 31% of the samples are shown at  $\geq 20$  and  $\geq 10$ , respectively. **c**, Contrasting allele frequencies for Strombolian and Plinian source loci (sample sizes shown under each axis label). The box denotes the interquartile range, with the median marked as a white point. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Hypothesis testing was performed using two-sided Mann-Whitney *U*-tests without correction for multiple tests. **d**, Numbers of active and hot source L1 elements per donor. Data are mean  $\pm$  s.d. number of elements per donor. **e**, The novel Plinian source element on 7p12.3 mediates 72 transductions among only 6 cancer samples. This generates a transduction that induces the deletion of the tumour-suppressor gene *CDKN2A*. **f**, Violin plots show the estimated number of distinct germline MEI alleles per PCAWG donor. The box denotes the interquartile range, with the median marked as a white point. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is less. Donors are grouped according to their genetic ancestry: AFR, African;

AMR, admixed American; EAS, East Asian; EUR, European; SAS, South Asian. Sample sizes are shown under each axis label. **g**, For each type of MEI (L1, Alu and SVA) identified both in PCAWG and in the 1000 Genomes Project (1KGP), the correlations between allele frequency estimates per ancestry derived from both projects are displayed in a blue (0) to red (1) coloured gradient.  $n = 2,583$  PCAWG patients. Two-sided hypothesis testing was performed using Spearman's rank correlations without correction for multiple tests. **h**, Example correlation between MEI allele frequencies derived from PCAWG and the 1000 Genomes Project for individuals with European ancestry ( $n = 1,201$  patients in PCAWG). Two-sided hypothesis testing was performed using Spearman's rank correlations without correction for multiple tests. **i**, Evaluation of TraFiC-mem false-discovery rate on a liver hepatocellular carcinoma sample (DOS0807) and a cell line (NCI-BL2087) sequenced using single-molecule sequencing with MinION (Oxford Nanopore). For each allele frequency bin (common,  $>5\%$ ; low frequency,  $1-5\%$ ; rare,  $<1\%$ ), the percentage of events supported by *N* long reads is represented (*N* ranges from 0–1 to more than 5). MEIs supported by at least two Nanopore reads were considered to be true positives (blue palette) and were classified as false positives (red) otherwise. The total number of germline MEIs per allele frequency bin is shown on the right. **j**, Correlation between predicted MEI lengths from Illumina and Nanopore data. Two-sided hypothesis testing was performed using Spearman's rank correlations without correction for multiple testing.



Extended Data Fig. 13 | See next page for caption.



**Extended Data Fig. 13 | Different mechanisms of telomere lengthening in cancer.** **a**, Scatter plot showing the four clusters of tumour-specific telomere patterns identified across PCAWG samples, together with the clusters of matched normal samples, generated by *t*-distributed stochastic neighbour embedding. Circles represent tumour samples and triangles represent matched normal samples. Points are coloured by tissue of origin. Data are based on  $n = 2,518$  tumour samples and their matched normal samples. **b**, Patterns of comutation of the relevant driver mutations across individual patients. Columns in plot represent individual patients, coloured by type of abnormality observed. **c**, Distribution of clonality of driver mutations in genes relevant to telomere maintenance across clusters. Clonal [early], clonal

mutations that occurred before duplications involving the relevant chromosome (including whole-genome duplications); clonal [late], clonal mutations that occurred after such duplications; and clonal [NA], mutations that occurred when no duplication was observed. **d**, Relationship between the estimated number of stem cell divisions per year and rate of telomere maintenance abnormalities across tumour types. The analysis uses data on estimated rates of stem cell division per year across  $n = 19$  tissue types previously collated from the literature<sup>82</sup>. Tumour types are coloured according to the scheme shown in Extended Data Fig. 3. Two-sided hypothesis testing was performed using likelihood ratio tests on Poisson regression models with no correction for multiple tests.

Extended Data Table 1 | Overview of the tumour types included in PCAWG project

Organ	Abbreviation	Included Subtypes	Cases	Sex		Age	
Neural Crest			Num.	F	M	Med.	10 <sup>th</sup> -90 <sup>th</sup>
CNS	CNS-GBM	Glioblastoma	41	13	28	60	43-72
CNS	CNS-Medullo	Medulloblastoma and variants	146	67	79	9	3-28
CNS	CNS-Oligo	Oligodendroglioma	18	9	9	41	21-62
CNS	CNS-PiloAstro	Pilocytic astrocytoma	89	47	42	8	2-17
Skin	Skin-Melanoma	Malignant melanoma	107	38	69	57	37-78
Endoderm							
Biliary	Biliary-AdenoCA	Papillary cholangiocarcinoma	34	15	19	64	53-76
Bladder	Bladder-TCC	Transitional cell carcinoma	23	8	15	65	52-80
Colon/Rectum	ColoRect-AdenoCA	Adenocarcinoma; Mucinous adeno.	60	30	30	67	46-81
Oesophagus	Eso-AdenoCA	Adenocarcinoma	98	14	84	70	56-79
Liver	Liver-HCC	Hepatocellular carcinoma; Comb. HCC/cholangio	317	89	228	67	50-78
Lung	Lung-AdenoCA	Adenocarcinoma; Adenocarcinoma <i>in situ</i>	38	20	18	66	47-77
Lung	Lung-SCC	Squamous cell carcinoma; Basaloid SCC	48	10	38	68	54-77
Pancreas	Panc-AdenoCA	Adeno.; Acinar cell Ca.; Mucinous adeno.	239	119	120	67	50-79
Pancreas	Panc-Endocrine	Neuroendocrine carcinoma	85	30	55	59	38-75
Prostate	Prost-AdenoCA	Adenocarcinoma	210	0	210	59	47-71
Stomach	Stomach-AdenoCA	Adenocarcinoma; Mucinous; Papillary; Tubular	75	18	57	65	47-79
Thyroid	Thy-AdenoCA	Adenocarcinoma; Columnar cell; Follicular type	48	37	11	51	26-75
Mesoderm							
Bone/Soft Tissue	Bone-Benign	Osteoblastoma; Osteofibrous dysplasia	7	4	3	18	12-30
Bone/Soft Tissue	Bone-Benign	Chondroblastoma; Chondromyxoid fibroma	9	2	7	16	14-38
Bone/Soft Tissue	Bone-Epith	Adamantinoma; Chordoma	10	4	6	60	37-67
Bone/Soft Tissue	Bone-Osteosarc	Osteosarcoma	38	20	18	20	9-58
Bone/Soft Tissue	SoftTissue-Leiomyo	Leiomyosarcoma	15	10	5	61	51-78
Bone/Soft Tissue	SoftTissue-Liposarc	Liposarcoma	19	5	14	n/a	n/a
Cervix	Cervix-AdenoCA	Adenocarcinoma	2	2	0	39	33-46
Cervix	Cervix-SCC	Squamous cell carcinoma	18	18	0	39	25-58
Head/Neck	Head-SCC	Squamous cell carcinoma	57	10	47	53	34-71
Kidney	Kidney-ChRCC	Adenocarcinoma, chromophobe type	45	19	26	47	34-69
Kidney	Kidney-RCC	Clear cell adenocarcinoma; papillary type	144	54	90	60	48-75
Lymphoid	Lymph-BNHL	Burkitt; Diffuse large B-cell; Follicular; Marginal	107	51	56	57	10-74
Lymphoid	Lymph-CLL	Chronic lymphocytic leukaemia	95	31	64	62	46-78
Myeloid	Myeloid-AML	Acute myeloid leukaemia	10	3	7	50	35-56
Myeloid	Myeloid-MDS	Myelodysplastic syndrome	2	1	1	76	74-77
Myeloid	Myeloid-MPN	Myeloproliferative neoplasm	26	14	12	56	38-75
Ovary	Ovary-AdenoCA	Adenocarcinoma; Serous cystadenocarcinoma	113	113	0	60	48-74
Uterus	Uterus-AdenoCA	Adeno., endometrioid; Serous cystadeno.	51	51	0	69	57-81
Ectoderm							
Breast	Breast-AdenoCA	Infiltrating duct carcinoma; Medullary; Mucinous	198	197	1	56	39-76
Breast	Breast-DCIS	Duct micropapillary carcinoma	3	3	0	55	43-60
Breast	Breast-LobularCA	Lobular carcinoma	13	13	0	53	42-69
Total			2658	1189	1469	59	21-76

Adeno., adenocarcinoma; Ca., carcinoma; Comb., combined; F, female; HCC, hepatocellular carcinoma; M, male; Med, median; 10–90th, 10–90th centiles; SCC, squamous cell carcinoma.

## Ethical Considerations of Genomic Cloud Computing

The PCAWG project represents the first large-scale use of distributed cloud computing in genomics. The project involved the movement of large quantities of personal health information across multiple legal jurisdictions and responsible use of this data by several hundred international researchers. Donor consents were written to explicitly allow for broad research use of the data and for international data sharing. PCAWG was granted permission by the leads of each of the tumour data providers to store, analyse and distribute the data on academic and/or commercial compute clouds.

To ensure that the PCAWG personal data were handled in a manner consistent with the donor consents, authorised representatives of each of the academic clouds and high-performance computing facilities signed a commitment not to access controlled tier data beyond the minimum needed to administer it. We negotiated similar contractual terms with commercial cloud partners. Prior to accessing the data, each PCAWG researcher was required to obtain local Institutional Review Board approval for their proposed analytic projects, and obtained controlled tier authorisation from dbGaP (National Center for Biotechnology Information) and the ICGC DACO (Centre of Genomics and Policy at McGill University). To handle the data securely, we encrypted it while in motion and at rest. We used a central authentication and digital token generating system to enforce a strong data access protocol that required researchers to provide their TCGA and/or ICGC credentials prior to accessing controlled tier data. No data breach or other compromise of donor confidentiality is known to have occurred over the course of the PCAWG project, despite its extensive use of cloud computing.

**Extended Data Table 3 | Scientific output using PCAWG data, in bite-size chunks**

Scientific area	Key findings	Citation
<b>Driver mutations</b>		
Discovery of non-coding drivers	<ul style="list-style-type: none"> <li>Estimated ~10-fold more coding than non-coding driver point mutations.</li> <li>Variation in point mutation density in non-coding regions influenced more by mutational processes than selection.</li> </ul>	4
Drivers by pathways and networks	<ul style="list-style-type: none"> <li>Both coding and non-coding alterations contribute to cancer pathways.</li> <li>Some pathways, such as RNA splicing, are primarily driven by non-coding mutations.</li> </ul>	16
<b>Evolution and heterogeneity</b>		
Timing of cancer evolution	<ul style="list-style-type: none"> <li>Each tumour type has a distinct pattern of early and late-occurring driver events.</li> <li>Earliest somatic mutations may occur decades prior to diagnosis, providing opportunities for early diagnosis.</li> <li>Intra-tumour heterogeneity is widespread and tumour subclones contain drivers that are under positive selection.</li> </ul>	7
<b>Structural variants</b>		
Patterns of structural variation	<ul style="list-style-type: none"> <li>Replication-based mechanisms of genome rearrangement frequent in many cancers, often causing driver structural variants.</li> <li>16 signatures of SV, including break-and-ligate patterns and copy-and-insert patterns, varying by size range, replication timing, tumour type and patient.</li> </ul>	6
Functional consequence of structural variation	<ul style="list-style-type: none"> <li>52 regions with recurrent structural breakpoints and 90 recurrently fused pairs of loci show evidence of positive selection.</li> <li>Oncogenic fusions are shaped by juxtaposition of proto-oncogenes with tissue-specific regulatory elements.</li> </ul>	4
Patterns of retrotransposition	<ul style="list-style-type: none"> <li>Many flavours of somatic retrotransposition in many cancers: LINE element mobilisation; transductions, pseudogenes, Alu elements.</li> <li>Retrotranspositions can induce genomic instability, including large deletions and breakage-fusion-bridge cycles amplifying cancer genes.</li> </ul>	10
Chromothripsis	<ul style="list-style-type: none"> <li>Chromothripsis pervasive across cancers, with frequency &gt;50% in several tumour types.</li> <li>Replicative processes and templated insertions contribute to rearrangement.</li> </ul>	18
<b>Mutational signatures</b>		
Signatures of point mutations	<ul style="list-style-type: none"> <li>&gt;70 distinct mutational signatures, encompassing SNVs, doublet subs and indels.</li> <li>Multiple signatures from unknown processes of DNA damage, repair and replication.</li> </ul>	5
Mutation distribution across genome	<ul style="list-style-type: none"> <li>Uneven distribution of somatic mutations and structural variants across the genome explained by epigenetic state of tissue, cell of origin and topological associated domains.</li> <li>Can be used to identify a tumour's type and presumed tissue/cell of origin.</li> </ul>	11,12,15
<b>Transcriptional consequences of somatic mutation</b>		
RNA effects of somatic mutation	<ul style="list-style-type: none"> <li>Genomic basis for RNA alterations across ~1200 tumours, including quantitative trait loci, allele specific expression and alternative splicing.</li> <li>Link between mutational signatures and expression; classification of gene fusions; identification of genes recurrently altered at RNA level.</li> </ul>	8,9
<b>Others</b>		
Tumour subtypes from genome sequencing	<ul style="list-style-type: none"> <li>Genomic distribution of somatic mutations, mutational signatures and driver mutations accurately distinguish major tumour types of primaries and metastases.</li> </ul>	12
Mitochondrial DNA mutations	<ul style="list-style-type: none"> <li>Somatic mitochondrial truncating mutations frequent in certain cancer types, associated with activation of critical signaling pathways.</li> </ul>	14
Telomere biology and sequences	<ul style="list-style-type: none"> <li>Activating <i>TERT</i> promoter mutations are the single most frequent non-coding driver.</li> <li>In <i>ATRX/DAXX</i>-mutant tumours, aberrant telomere variant repeat distribution is common.</li> </ul>	4,13

Key findings are described further in associated papers<sup>4-18</sup>.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

Data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at <https://www.overture.bio>. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to [contact@overture.bio](mailto:contact@overture.bio).

#### Data analysis

The workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Docker images and available at: <https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files>. Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACESeq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBba v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15; Chromothripsis Explorer v1.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

WGS somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA



projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014. No statistical methods were used to predetermine sample size.
Data exclusions	After quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine). Exclusion criteria were pre-determined.
Replication	In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement.
Randomization	No randomisation was performed - this was a descriptive study, not an experimental study.
Blinding	No blinding was undertaken - this was a descriptive study, not an experimental study.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-O-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour
----------------------------	--

types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.

## Recruitment

Patients were recruited by the participating centres following local protocols. Samples obtained had to meet criteria on amount of tumour DNA available, meaning that the cohort is potentially somewhat biased towards larger tumours. Otherwise, we anticipate no major recruitment biases.

## Ethics oversight

The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# The repertoire of mutational signatures in human cancer

<https://doi.org/10.1038/s41586-020-1943-3>

Received: 18 May 2018

Accepted: 18 November 2019

Published online: 5 February 2020

Open access

Ludmil B. Alexandrov<sup>1,25</sup>, Jaegil Kim<sup>2,25</sup>, Nicholas J. Haradhvala<sup>2,3,25</sup>, Mi Ni Huang<sup>4,5,25</sup>, Alvin Wei Tian Ng<sup>4,5</sup>, Yang Wu<sup>4,5</sup>, Arnoud Boot<sup>4,5</sup>, Kyle R. Covington<sup>6,7</sup>, Dmitry A. Gordenin<sup>8</sup>, Erik N. Bergstrom<sup>1</sup>, S. M. Ashiqul Islam<sup>1</sup>, Nuria Lopez-Bigas<sup>9,10,11</sup>, Leszek J. Klimczak<sup>12</sup>, John R. McPherson<sup>4,5</sup>, Sandro Morganello<sup>13</sup>, Radhakrishnan Sabarinathan<sup>10,14,15</sup>, David A. Wheeler<sup>6,16</sup>, Ville Mustonen<sup>17,18,19</sup>, PCAWG Mutational Signatures Working Group<sup>20</sup>, Gad Getz<sup>2,3,21,22,26</sup>, Steven G. Rozen<sup>4,5,23,26\*</sup>, Michael R. Stratton<sup>13,26\*</sup> & PCAWG Consortium<sup>24</sup>

Somatic mutations in cancer genomes are caused by multiple mutational processes, each of which generates a characteristic mutational signature<sup>1</sup>. Here, as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium<sup>2</sup> of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), we characterized mutational signatures using 84,729,690 somatic mutations from 4,645 whole-genome and 19,184 exome sequences that encompass most types of cancer. We identified 49 single-base-substitution, 11 doublet-base-substitution, 4 clustered-base-substitution and 17 small insertion-and-deletion signatures. The substantial size of our dataset, compared with previous analyses<sup>3–15</sup>, enabled the discovery of new signatures, the separation of overlapping signatures and the decomposition of signatures into components that may represent associated—but distinct—DNA damage, repair and/or replication mechanisms. By estimating the contribution of each signature to the mutational catalogues of individual cancer genomes, we revealed associations of signatures to exogenous or endogenous exposures, as well as to defective DNA-maintenance processes. However, many signatures are of unknown cause. This analysis provides a systematic perspective on the repertoire of mutational processes that contribute to the development of human cancer.

Somatic mutations in cancer genomes are caused by mutational processes of both exogenous and endogenous origin that operate during the cell lineage between the fertilized egg and the cancer cell<sup>16</sup>. Each mutational process may involve components of DNA damage or modification, DNA repair and DNA replication (which may be normal or abnormal), and generates a characteristic mutational signature that potentially includes base substitutions, small insertions and deletions (indels), genome rearrangements and chromosome copy-number changes<sup>1</sup>. The mutations in an individual cancer genome may have been generated by multiple mutational processes, and thus incorporate multiple superimposed mutational signatures. Therefore, to systematically characterize the mutational processes that contribute to

cancer, mathematical methods have previously been used to decipher mutational signatures from somatic mutation catalogues, estimate the number of mutations that are attributable to each signature in individual samples and annotate each mutation class in each tumour with the probability that it arose from each signature<sup>6,9,17–27</sup>.

Previous studies of multiple types of cancer have identified more than 30 single-base substitution (SBS) signatures, some of known—but many of unknown—aetiologies, some ubiquitous and others rare, some part of normal cell biology and others associated with abnormal exposures or neoplastic progression<sup>3–5,7–15</sup>. Genome rearrangement signatures have also previously been described<sup>11,25,28–30</sup>. However, the analysis of other classes of mutation has been relatively limited<sup>3,11,31–33</sup>.

<sup>1</sup>Department of Cellular and Molecular Medicine, Department of Bioengineering, Moores Cancer Center, University of California, San Diego, CA, USA. <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>3</sup>Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. <sup>4</sup>Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore. <sup>5</sup>Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore. <sup>6</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. <sup>7</sup>Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA. <sup>8</sup>Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. <sup>9</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>10</sup>Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain. <sup>11</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. <sup>12</sup>Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. <sup>13</sup>Wellcome Sanger Institute, Hinxton, UK. <sup>14</sup>National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India. <sup>15</sup>Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. <sup>16</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. <sup>17</sup>Department of Computer Science, University of Helsinki, Helsinki, Finland. <sup>18</sup>Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland. <sup>19</sup>Institute of Biotechnology, University of Helsinki, Helsinki, Finland. <sup>20</sup>A list of members and their affiliations appears at the end of the paper. <sup>21</sup>Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. <sup>22</sup>Harvard Medical School, Boston, MA, USA. <sup>23</sup>SingHealth, Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore, Singapore. <sup>24</sup>A list of members and their affiliations appears in the Supplementary Information. <sup>25</sup>These authors contributed equally: Ludmil B. Alexandrov, Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang. <sup>26</sup>These authors jointly supervised this work: Gad Getz, Steven G. Rozen, Michael R. Stratton. \*e-mail: [steverozen@gmail.com](mailto:steverozen@gmail.com); [mrs@sanger.ac.uk](mailto:mrs@sanger.ac.uk)

Mutational signature analysis has predominantly used cancer exome sequences. However, the many-fold-greater numbers of somatic mutations in whole genomes provide substantially increased power for signature decomposition, enabling the better separation of partially correlated signatures and the extraction of signatures that contribute relatively small numbers of mutations. Furthermore, technical artefacts and differences in sequencing technologies and mutation-calling algorithms can themselves generate mutational signatures. Therefore, the uniformly processed and highly curated sets of all classes of somatic mutations from the 2,780 cancer genomes of the PCAWG project<sup>2</sup>, combined with most other suitable cancer genomes (accession code syn11801889, available at <https://www.synapse.org/#!Synapse:syn11801889>), present a notable opportunity to establish the repertoire of mutational signatures and determine their activities across different types of cancer. The timing of these signatures during the evolution of individual cancers and the repertoire of signatures of structural variation have been explored in other PCAWG analyses<sup>30,34</sup>.

## Mutational signature analysis

The 23,829 samples—which include most types of cancer, and comprise the 2,780 PCAWG whole genomes<sup>2</sup>, 1,865 additional whole genomes and 19,184 exomes—yielded 79,793,266 somatic SBSs, 814,191 doublet-base substitutions (DBSs) and 4,122,233 small indels that were analysed for mutational signatures, about 10-fold-more mutations than any previous study of which we are aware (syn11801889)<sup>6</sup>.

We developed classifications for each type of mutation. For SBSs, the primary classification comprised 96 classes (available at <https://cancer.sanger.ac.uk/cosmic/signatures/SBS>) constituted by the 6 base substitutions C>A, C>G, C>T, T>A, T>C and T>G (in which the mutated base is represented by the pyrimidine of the base pair), plus the flanking 5' and 3' bases. In some analyses, two flanking bases 5' and 3' to the mutated base were considered (producing 1,536 classes) or mutations within transcribed genome regions were selected and classified according to whether the mutated pyrimidine fell on the transcribed or untranscribed strand (producing 192 classes). We also derived a classification for DBSs (78 classes; available at <https://cancer.sanger.ac.uk/cosmic/signatures/DBS>). Indels were classified as deletions or insertions and—when of a single base—as C or T, and according to the length of the mononucleotide repeat tract in which they occurred. Longer indels were classified as occurring at repeats or with overlapping microhomology at deletion boundaries, and according to the size of indel, repeat and microhomology (83 classes; available at <https://cancer.sanger.ac.uk/cosmic/signatures/ID>).

The PCAWG whole-genome sequences, the additional whole-genome sequences and the exome sequences were each analysed separately (syn11801889)<sup>2</sup>. Signatures were extracted from each type of cancer individually, from all cancer types together, as separate SBS, DBS and indel signatures, and as composite signatures of all three types of mutation (Supplementary Note 2).

We used two methods based on nonnegative matrix factorization (NMF): SigProfiler, an elaborated version of the framework used for the previous 'Catalogue Of Somatic Mutations In Cancer' (COSMIC) compendium of mutational signatures (COSMIC v.2, available at [https://cancer.sanger.ac.uk/cosmic/signatures\\_v2](https://cancer.sanger.ac.uk/cosmic/signatures_v2))<sup>11,17</sup>, and SignatureAnalyzer, which is based on a Bayesian variant of NMF<sup>9,27,35</sup>. NMF determines the signature profiles and contributions of each signature to each cancer genome as part of its factorization of the input matrix of mutation spectra. However, with many signatures and/or heterogeneous mutation burdens across samples, the mutations observed in a particular sample can be reconstructed in multiple ways—often with small and/or biologically implausible contributions from many signatures. Therefore, each method has developed a separate procedure for estimating the contributions of signatures to each sample (Methods).

We tested SignatureAnalyzer and SigProfiler on 11 sets of synthetic data (including 64,400 synthetic samples), generated from known

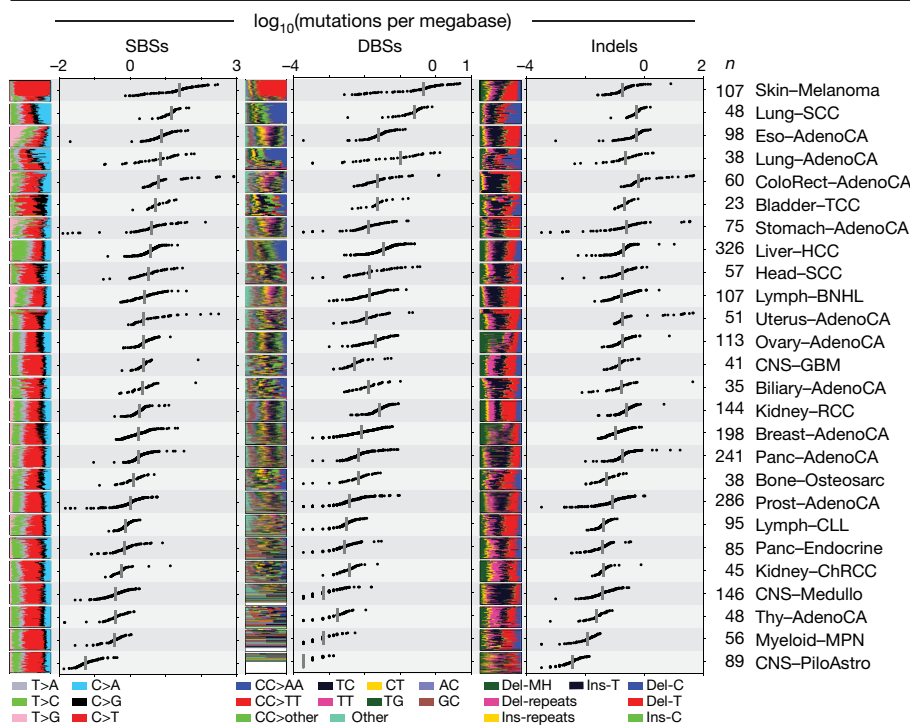
signature profiles (Methods, Supplementary Note 2). Both methods performed well in re-extracting known signatures from realistically complex data. Extracted signatures that were discordant from the known input usually arose from difficulties in selecting the correct number of signatures. The results confirm that use of NMF-based approaches for extracting mutational signatures is not a purely algorithmic process, but also requires consideration of evidence from experimentally determined mutational signatures and the DNA damage and repair literature, prior evidence of biological plausibility and human-guided sensitivity analysis confirming that extractions from different groupings of tumours yield consistent results. We used these types of evidence and approaches in determining the signature profiles reported here. The findings are consistent with results regarding NMF, and the related areas of probabilistic topic modelling and latent Dirichlet allocation, in multiple problem domains<sup>36,37</sup>. It is widely understood that the choice of the number of latent variables (for our purposes, the number of mutational signatures) is rarely amenable to complete automation.

The results from our SigProfiler and SignatureAnalyzer analyses of cancer data exhibited many similarities, and we assigned the same identifiers to similar signatures extracted using the two methods (syn12016215). However, there were also noteworthy differences. The numbers of SBS signatures found in PCAWG tumours with a low mutation burden (94.4% of cases that contain 47% of mutations) were similar: 31 using SigProfiler and 35 using SignatureAnalyzer. However, the numbers of additional SBS signatures extracted from hypermutated PCAWG samples (5.6% of cases, containing 53% of mutations) were different: 13 using SigProfiler and 25 using SignatureAnalyzer. There were also differences in SBS signature profiles, including among signatures found in cases with a low mutation burden. The latter primarily involved relatively featureless ('flat') signatures, which are mathematically challenging to deconvolute. Finally, there were differences in signature attributions to individual samples. SignatureAnalyzer used more signatures to reconstruct the mutational profiles (Extended Data Fig. 1) (syn12169204 and syn12177011) and attributions to flat signatures were different (Extended Data Fig. 2a, b) (syn12169204). The DBS and indel signatures were generally similar between the two methods (Extended Data Fig. 2c, d).

The final reference mutational signatures were determined from the PCAWG set, supplemented by additional signatures from the other datasets (COSMIC, available at <https://cancer.sanger.ac.uk/cosmic/signatures>). Each signature was allocated an identifier consistent with, and extending, the COSMIC v.2 annotation. Some previous signatures split into multiple constituent signatures: these were numbered as in the previous annotation, but with additional letter suffixes (for example, SBS17 was split into SBS17a and SBS17b). DNA sequencing and analysis artefacts also generate mutational signatures. We indicate which signatures are possible artefacts but do not present them below (full information is available at <https://cancer.sanger.ac.uk/cosmic/signatures>). The results of both SignatureAnalyzer and SigProfiler were used throughout the study. However, for brevity and for continuity with the signature set previously displayed in COSMIC v.2—which has been widely used as a reference—SigProfiler results are outlined here, and SignatureAnalyzer results are provided in Extended Data Figs. 3, 4 and at syn11738307.

## Single-base substitution signatures

There were substantial differences in the numbers of SBSs between samples (ranging from hundreds to millions) and between cancer types<sup>38</sup> (Fig. 1). In total, 67 SBS mutational signatures were extracted, of which 49 were considered likely to be of biological origin (Fig. 2, Methods; available at <https://cancer.sanger.ac.uk/cosmic/signatures/SBS/>). Except for signature SBS25, all signatures reported in COSMIC v.2 (ref. <sup>6</sup>) were confirmed; the median cosine similarity between the newly



**Fig. 1 | Mutation burdens of SBSs, DBSs and small indels across PCAWG tumour types.** The numbers of cases of each tumour type are shown next to the labels. Each dot represents one tumour. Tumour types are ordered by the median numbers of single-base substitutions. Only tumour types with >20 samples are shown. AdenoCA, adenocarcinoma; BNHL, B-cell non-Hodgkin lymphoma; ChRCC, chromophobe renal cell carcinoma; CLL, chronic lymphocytic leukaemia; CNS, central nervous system; ColoRect, colorectal; Eso, oesophageal; GBM, glioblastoma; HCC, hepatocellular carcinoma; Medullo, medulloblastoma; MH, microhomology; MPN, myeloproliferative neoplasm; Osteosarc, osteosarcoma; Panc, pancreatic; PiloAstro, pilocytic astrocytoma; Prost, prostate; RCC, renal cell carcinoma; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; Thy, thyroid.

derived signatures and those on COSMIC v.2 was 0.95, excluding the 'split' signatures (discussed below). SBS25 was previously found in cell lines derived from Hodgkin lymphomas treated with chemotherapy, and no primary cancers of this type were available. The newly derived signatures showed much improved separation from each other and more-distinct signature profiles, as compared with COSMIC v.2 signatures (see 'Better separation compared to COSMIC v.2 signatures' in Supplementary Note 2 for more information).

Thirteen of the SBS signatures we extracted (excluding those due to signature splitting) represent newly identified and probably real signatures, not present in COSMIC v.2. Some were rare (SBS31, SBS32, SBS35, SBS36, SBS42 and SBS44). Others were more common, but contributed relatively few mutations and/or were similar to previously discovered signatures (SBS38, SBS39 and SBS40). Notably, SBS40 is a flat signature similar to SBS5. It contributes to multiple types of cancer, but its similarity to SBS5 renders the extent of this contribution uncertain. For some of the newly identified signatures, there were plausible underlying aetiologies (Fig. 3, Extended Data Figs. 4, 5): for SBS31 and SBS35, platinum compound chemotherapy<sup>39</sup>; for SBS32, azathioprine therapy; for SBS36, inactivating germline or somatic mutations in *MUTYH* (which encodes a component of the base excision repair machinery)<sup>40,41</sup>; for SBS38, additional effects of exposure to ultraviolet (UV) light; for SBS42, occupational exposure to haloalkanes<sup>13</sup>; and for SBS44, defective DNA mismatch repair<sup>42</sup>.

Three previously characterized base substitution signatures (SBS7, SBS10 and SBS17) split into multiple constituent signatures (Fig. 2). Signature splitting probably reflects the existence of multiple distinct mutational processes initiated by the same exposure that have closely—but not perfectly—correlated activities. We previously regarded SBS7 as a single signature composed predominantly of C>T at CCN and TCN trinucleotides (the mutated base is underlined) together with many fewer T>N mutations. It was found in malignant melanomas and squamous skin carcinomas, and is probably due to the UV-light-induced formation of pyrimidine dimers, followed by translesion DNA synthesis by error-prone polymerases predominantly inserting A opposite to damaged cytosines. SBS7 has now been decomposed into four constituent signatures. SBS7a and SBS7b (consisting mainly of C>T at TCN and C>T at CCN, respectively) may reflect different pyrimidine-dimer

photoproducts. SBS7c and SBS7d (consisting predominantly of T>A at NIT and T>C at NIT, respectively<sup>43</sup>) may be due to low frequencies of the misincorporation of T and G opposite to thymines in pyrimidine dimers. The splitting of SBS10 and SBS17 is described at <https://cancer.sanger.ac.uk/cosmic/signatures/SBS/>.

Several base substitution signatures showed transcriptional strand bias, which may be attributable to transcription-coupled nucleotide excision repair acting on DNA damage and/or to an excess of DNA damage on untranscribed strands of genes<sup>44</sup>. Both mechanisms result in more mutations of damaged bases on untranscribed than on transcribed strands of genes. Assuming that either mechanism is responsible for the observed transcriptional strand biases, DNA damage to cytosine (SBS7a and SBS7b), guanine (SBS4, SBS8, SBS19, SBS23, SBS24, SBS31, SBS32, SBS35 and SBS42), thymine (SBS7c, SBS7d, SBS21, SBS26 and SBS33) and adenine (SBS5, SBS12, SBS16, SBS22 and SBS25) may underlie these mutational signatures (plots of strand bias are available at <https://cancer.sanger.ac.uk/cosmic/signatures/SBS/>). The likely DNA-damaging agents are known for SBS4 (tobacco mutagens), SBS7a, SBS7b, SBS7c and SBS7d (UV light), SBS22 (aristolochic acid), SBS24 (aflatoxin), SBS25 (chemotherapy), SBS31 and SBS35 (platinum compounds), SBS32 (azathioprine) and SBS42 (haloalkanes).

Using the SBS classification of 1,536 mutation types, which uses the sequence context two bases 5' and two bases 3' to each mutated base, yielded signatures that are largely consistent with those based on substitutions in trinucleotide contexts. Notably, however, two forms of both SBS2 and SBS13 were extracted, one with mainly a pyrimidine and the other with mainly a purine at the −2 base (the second base 5' to the mutated cytosine). These may represent the activities of the cytidine deaminases APOBEC3A and APOBEC3B, respectively<sup>45</sup>. If so, APOBEC3A accounts for many more mutations than APOBEC3B in cancers with high APOBEC activity. Other signatures showed nonrandom sequence contexts at +2 and −2 positions (for example, SBS17a, SBS17b and SBS9), but sequence context effects were generally much stronger for bases immediately 5' and 3' to mutated bases.

SBS signatures showed substantial variation in the numbers of cancer types and cancer samples in which they were found, and in the mutations attributed per cancer sample (Fig. 3). Almost all individual cancer samples exhibited multiple signatures, with a mode of three in



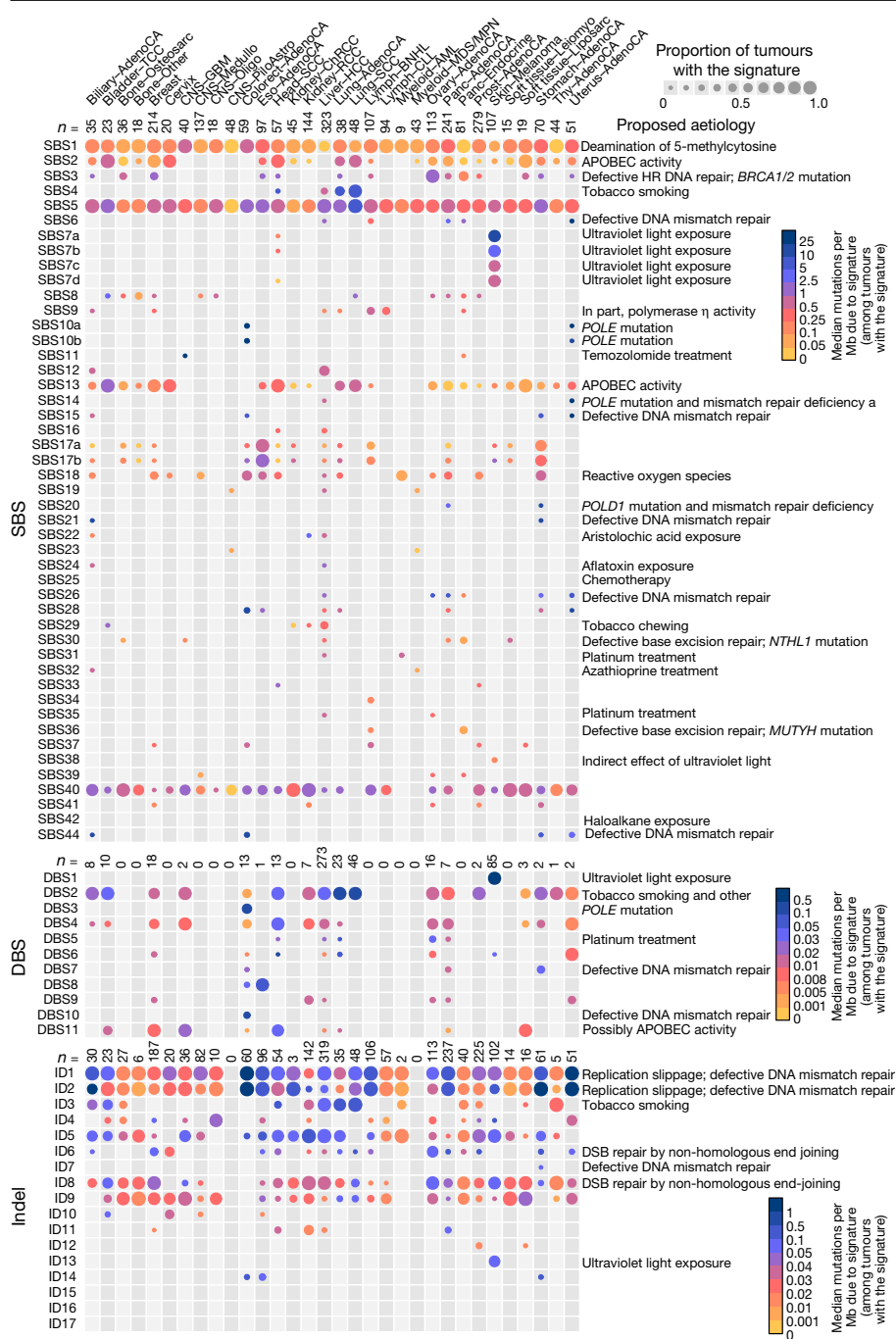


**Fig. 2 | Profiles of SBS, DBS and small indel mutational signatures.** The classifications of each mutation type (SBS, 96 classes; DBS, 78 classes; and indels, 83 classes) are described in the main text. Magnified versions of signatures SBS4, DBS2 and ID3 (all of which are associated with tobacco

smoking) are shown to illustrate the positions of each mutation subtype on each plot. The plotted data are available in digital form (along with the x axis labels) at syn12025148.

the PCAWG set (syn12169204). The assigned signatures reconstruct well the mutational spectra of the cancer samples (in PCAWG samples, the median cosine similarity was 0.97; 96.3% of samples with cosine similarity >0.90): Fig. 4 shows illustrative examples.

Some mutational processes generate base substitutions that cluster in small genomic regions. The limited numbers of such mutations may result in a failure to detect their signatures using standard methods. We therefore identified clustered mutations in each genome and analysed



**Fig. 3 | The number of mutations contributed by each mutational signature to the PCAWG tumours.** The size of each dot represents the proportion of samples of each tumour type that shows the mutational signature. The colour of each dot represents the median mutation burden of the signature in samples that show the signature. Tumours that had few mutations or that were poorly reconstructed by the signature assignment were excluded. The contributions of composite signatures to the PCAWG cancers, and SBS signatures to the complete set of cancer samples analysed, are shown in Extended Data Figs. 4 and 5, respectively. AML, acute myeloid leukaemia; liposarc, liposarcoma; MDS, myelodysplastic syndrome.

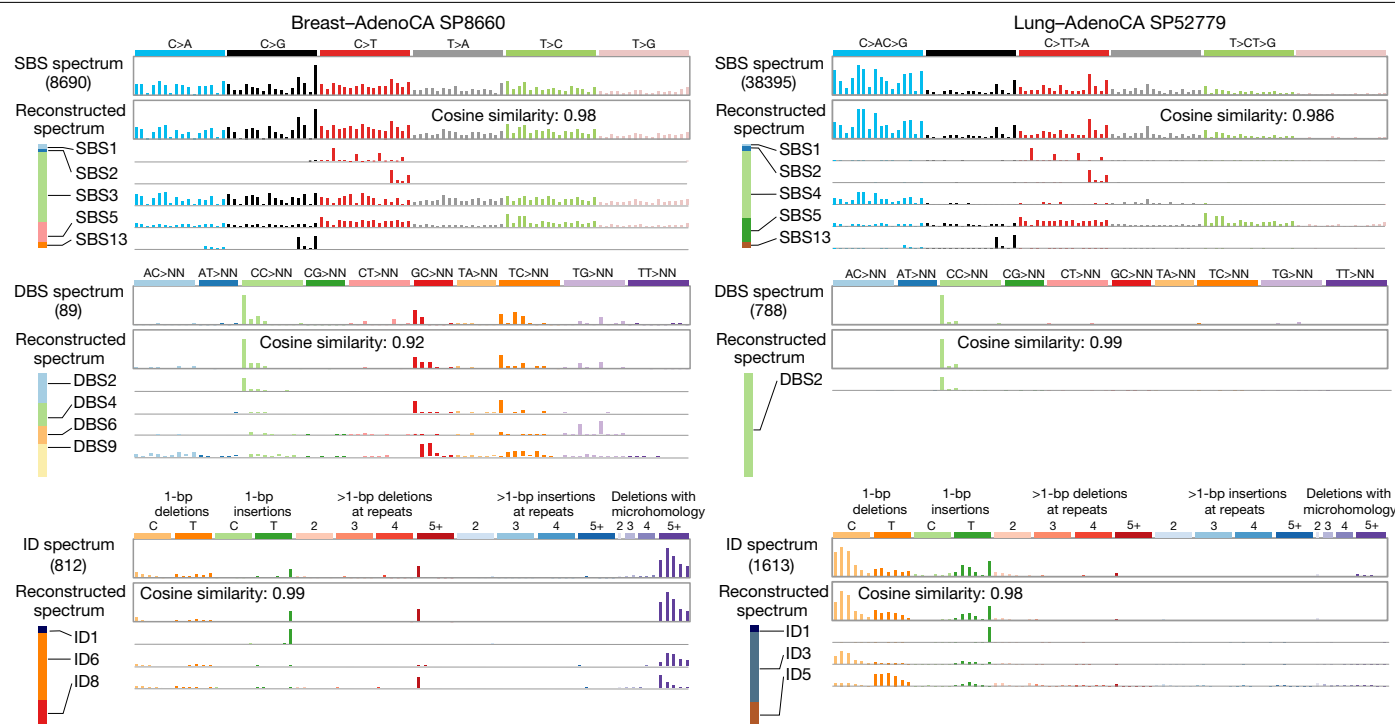
them separately (Methods). Four main clustered mutational signatures were identified (Fig. 2), as previously reported<sup>4,27,32</sup>. Two, which are found in multiple types of cancer, were similar to SBS2 and SBS13 (which have been attributed to APOBEC enzyme activity) and represent foci of kataegis<sup>3,32,46</sup>. Two further clustered signatures, one characterized by C>T and C>G mutations at (A or G)C(C or T) trinucleotides<sup>47</sup> and the other T>A and T>C mutations at (A or T)I(A or T), were found in lymphoid neoplasms; they probably represent the direct and indirect consequences of activation-induced cytidine deaminase mutagenesis and translesion DNA synthesis by error-prone polymerases (SBS84 and SBS85, respectively)<sup>27</sup>.

### Doublet-base substitution signatures

Tandem doublet, triplet, quadruplet, quintuplet and sextuplet base substitutions (syn11801938 and syn11726620) were observed at about 1% the prevalence of SBSs. In most cancer genomes, the number of DBS

was considerably higher than would be expected from the random adjacency of SBSs (syn12177057), indicating the existence of commonly occurring, single mutagenic events that cause substitutions at neighbouring bases. There was substantial variation in the number of DBSs, ranging from 0 to 20,818 in a sample. The numbers of DBSs were generally proportional to the numbers of SBSs (Fig. 1), although colorectal adenocarcinomas had fewer than expected, and lung cancers and melanomas had more (Extended Data Table 1). We extracted eleven DBS signatures (Fig. 2, of which three have previously been reported<sup>33,48</sup>).

Signature DBS1 was characterized by CC>TT mutations (Fig. 2), contributed hundreds to tens of thousands of mutations in malignant melanomas with SBS7a and SBS7b (Fig. 3), exhibited transcriptional strand bias consistent with damage to cytosines (syn12177063) and is a known consequence of DNA damage induced by UV light<sup>33,49</sup>. Excluding cancers associated with exposure to UV light also yielded a signature (DBS11) that was characterized predominantly by CC>TT mutations, but only



**Fig. 4 | Illustrative examples of mutational spectra of individual cancer samples.** The contributory SBS, DBS and small indel mutational signatures in two tumours are shown.

contributing tens of mutations in many samples from multiple types of cancer (Figs. 2, 3). DBS11 was associated with SBS2, which is due to APOBEC activity: APOBEC activity may, therefore, also generate DBS11.

DBS2 was composed predominantly of CC>AA mutations, with smaller numbers of CC>AG and CC>AT mutations, and contributed hundreds to thousands of mutations in lung adenocarcinoma, lung squamous and head and neck squamous carcinomas, which are often caused by tobacco smoking<sup>33</sup> (Figs. 2, 3). DBS2 showed transcriptional strand bias indicative of guanine damage (syn12177064) and was associated with SBS4, which is caused by exposure to tobacco smoke. It is likely, therefore, that DBS2 can be a consequence of DNA damage by tobacco-smoke mutagens.

A signature similar to DBS2 contributed hundreds of mutations to liver cancers and tens of mutations to other types of cancer without evidence of exposure to tobacco smoke. A pattern resembling DBS2 also dominates DBSs in healthy mouse cells<sup>50</sup>. The nature of the mutational processes that underlie these signatures in human cancers that are unrelated to smoking, and in healthy mice, is unknown. However, in experimental systems, acetaldehyde exposure has been shown to generate a mutational signature characterized primarily by CC>AA mutations, and lower burdens of CC>AG and CC>AT mutations, together with C>A SBSs<sup>48</sup>. Acetaldehyde is an oxidation product of alcohol and a constituent of cigarette smoke. The role of acetaldehyde, and perhaps other aldehydes, in generating DBS2 merits further investigation<sup>51</sup>.

DBS3, DBS7, DBS8 and DBS10 showed hundreds to thousands of mutations in rare colorectal, stomach and oesophageal cancers, some of which showed evidence of defective DNA mismatch repair (DBS7 and DBS10) or polymerase epsilon exonuclease domain mutations (DBS3) that generate hypermutator phenotypes (Figs. 2, 3). DBS5 was found in cancers exposed to platinum chemotherapy, and is associated with SBS31 and SBS35.

### Small insertion-and-deletion signatures

Indels were usually present at about 10% of the frequency of base substitutions (Fig. 1). There was substantial variation between cancer

genomes in the number of indels, even when cancers with evidence of defective DNA mismatch repair were excluded. Overall, the numbers of deletions and insertions were similar, but there was variation between cancer types: some cancers showed more deletions and others more insertions of various subtypes (Fig. 1). We extracted 17 indel mutational signatures (Fig. 2).

Indel signature 1 (ID1) was composed predominantly of insertions of thymine and ID2 was composed predominantly of deletions of thymine, both at long ( $\geq 5$ ) thymine mononucleotide repeats (Fig. 2). Tens to hundreds of mutations of both signatures were found in most samples of most types of cancer, but were particularly common in colorectal, stomach, endometrial and oesophageal cancers and in diffuse large B cell lymphoma (Fig. 3). Together, ID1 and ID2 accounted for 97% and 45% of indels in hypermutated and non-hypermutated cancer genomes, respectively (Extended Data Table 2). They are probably due to slippage of either the nascent (ID1) or template strand (ID2) during DNA replication of long mononucleotide tracts.

ID3 was characterized predominantly by deletions of cytosine at short ( $\leq 5$ -bp long) mononucleotide cytosine repeats and exhibited hundreds of mutations in cancers of the lung, head and neck that are associated with tobacco smoking (Figs. 2, 3). There was transcriptional strand bias of mutations, with more guanine deletions than cytosine deletions on the untranscribed strands of genes, which is compatible with transcription-coupled nucleotide excision repair of damaged guanine (syn12177065 and syn12177066). The numbers of ID3 mutations positively correlated with the numbers of SBS4 and DBS2 mutations, which we have shown are associated with tobacco smoking (Extended Data Figs. 6, 7). Thus, DNA damage by components of tobacco smoke probably underlie ID3.

ID13 was characterized predominantly by deletions of thymine at thymine–thymine dinucleotides and exhibited large numbers of mutations in malignant melanomas of the skin (Figs. 2, 3). The numbers of ID13 mutations correlated with the numbers of SBS7a, SBS7b and DBS1 mutations, which we have attributed to DNA damage induced by UV light (Extended Data Figs. 6, 7). However, deletions of cytosine

at cytosine–cytosine dinucleotides did not feature strongly in ID13, which may reflect the predominance of thymine compared to cytosine dimers induced by UV light<sup>52</sup>.

ID6 and ID8 were both characterized predominantly by  $\geq 5$ -bp deletions (Fig. 2). ID6 exhibited overlapping microhomology at deletion boundaries with a mode of 2 bp (and often longer stretches) and correlated with SBS3, which we have attributed to defective homologous-recombination-based repair (Extended Data Figs. 6, 7). By contrast, ID8 deletions showed shorter or no microhomology at deletion boundaries and did not strongly correlate with SBS3. Both deletion patterns may be characteristic of DNA double-strand-break repair by non-homologous-recombination-based end-joining mechanisms and—if so—this suggests that at least two distinct forms are operative in human cancer<sup>53</sup>.

A small fraction of cancers exhibited very large numbers of ID1 and ID2 mutations ( $>10,000$ ) (Fig. 3) (shown at <https://cancer.sanger.ac.uk/cosmic/signatures/ID>). These were usually accompanied by SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and/or SBS44, which are associated with deficiency in DNA mismatch repair—sometimes combined with POLE or POLD1 proofreading deficiency (SBS14 and SBS20)<sup>35</sup>. Occasional cases with these signatures additionally showed large numbers of indels attributed to ID7 (syn11738668), and rare samples showed large numbers of ID4, ID11, ID14, ID15, ID16 or ID17 mutations but did not show large numbers of ID1 and ID2 mutations or the SBS signatures associated with deficiency in DNA mismatch repair.

## Correlations with age

A positive correlation between age of cancer diagnosis and the number of mutations attributable to a signature suggests that the underlying mutational process has been operative (at a more or less constant rate) throughout the cell lineage from fertilized egg to cancer cell, and thus in the normal cells from which that type of cancer develops<sup>6,54</sup>. Confirming previous reports<sup>6,54</sup>, the numbers of SBS1 and SBS5 mutations correlate with age, and exhibit different rates in different types of tissue ( $q$  values provided in syn12030687, syn20317940 and syn12217988). SBS40 also correlated with age in multiple types of cancer, although—given its similarity to SBS5—misattribution cannot be excluded. DBS2 and DBS4 correlated with age; consistent with activity in normal cells and, when combined their profiles closely resemble the spectrum of DBS mutations found in normal mouse cells<sup>50</sup>. ID1, ID2, ID5 and ID8 showed correlations with age in multiple tissues. ID1 and ID2 indels are probably due to slippage at poly T repeats during DNA replication and correlated with the numbers of SBS1 substitutions, which have previously been proposed to reflect the number of mitoses a cell has experienced<sup>6</sup>. Thus, SBS1, ID1 and ID2 may all be generated during DNA replication at mitosis. The number of ID5 mutations correlated with the number of SBS40 mutations, and the mutational processes that underlie these two age-correlated signatures may therefore contain common components. ID8, which is predominantly composed of  $\geq 5$ -bp deletions with no or 1 bp of microhomology at their boundaries, is probably due to DNA double-strand breaks repaired by a non-homologous-end-joining mechanism. The results indicate that multiple mutational processes operate in normal cells.

## Discussion

There are important constraints, limitations and assumptions in the analytic frameworks used here to characterize mutational signatures. Signatures extracted from sample sets in which multiple processes are operative remain mathematical approximations, with profiles that are potentially influenced by the mathematical approach used and other factors. For conceptual and practical simplicity, we assume that a single signature is associated with each mutational process and provide an average reference signature to represent it. However, we do not discount the possibility that further nuances and variations

of signature profiles exist. We have estimated the contributions from each signature to the mutation burden in each sample. However, with increasing numbers of signatures and differences of multiple orders of magnitude in mutation burdens between some signatures, prior knowledge has helped to avoid biologically implausible results. Thus, the further development of methods for deciphering and attributing mutational signatures is warranted, ideally supported by signatures derived from experimental systems in which the causes are known. Nevertheless, signatures with many similarities and some differences can be found by different mathematical approaches, and these can be confirmed in several ways, including experimentally elucidated signatures<sup>5,31,39,42,43,54–62</sup> and tumours dominated by a single signature (syn12016215).

This analysis includes most publicly available exome and whole-genome cancer sequences. Some rare or geographically restricted signatures may not have been captured, signatures conferring limited mutation burdens may have been missed and signatures of therapeutic mutagenic exposures have not been exhaustively explored. Nevertheless, it is likely that a substantial proportion of the naturally occurring mutational signatures found in human cancer have now been described. This comprehensive repertoire provides a foundation for research into the aetiologies of geographical and temporal differences in cancer incidence, the mutational processes that operate in healthy tissues and non-neoplastic disease states, clinical and public health applications of signatures and mechanistic understanding of the mutational processes that underlie carcinogenesis.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1943-3>.

- Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
- Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
- Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
- Poon, S. L. et al. Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med.* **7**, 38 (2015).
- Schulze, K. et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
- Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
- Merlevede, J. et al. Mutation allele burden remains unchanged in chronic myelomonocytic leukaemia responding to hypomethylating agents. *Nat. Commun.* **7**, 10767 (2016).
- Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
- Petljak, M. & Alexandrov, L. B. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**, 531–540 (2016).
- Mimaki, S. et al. Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. *Carcinogenesis* **37**, 817–826 (2016).
- Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
- Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).
- Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
- Morganella, S. et al. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).

19. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, R39 (2013).
20. Roberts, N. *hdp (hierarchical Dirichlet process) R package* <https://github.com/nicolaroberts/hdp> (2015).
21. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
22. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.* **11**, e1005657 (2015).
23. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16 (2017).
24. Ardin, M. et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17**, 170 (2016).
25. Funnell, T. et al. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol.* **15**, e1006799 (2019).
26. Blokzijl, F., Janssen, R., van Bostel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
27. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
28. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
29. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
30. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* <https://doi.org/10.1038/s41586-019-1913-9> (2020).
31. Meier, B. et al. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.* **24**, 1624–1636 (2014).
32. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
33. Chen, J. M., Férec, C. & Cooper, D. N. Patterns and mutational signatures of tandem base substitutions causing human inherited disease. *Hum. Mutat.* **34**, 1119–1130 (2013).
34. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* <https://doi.org/10.1038/s41586-019-1907-7> (2020).
35. Haradhvala, N. J. et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018).
36. Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S.-i. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation* (John Wiley & Sons, 2009).
37. Blei, D., Carin, L. & Dunson, D. Probabilistic topic models: a focus on graphical model design and applications to document and image analysis. *IEEE Signal Process. Mag.* **27**, 55–65 (2010).
38. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
39. Boot, A. et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* **28**, 654–665 (2018).
40. Viel, A. et al. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).
41. Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).
42. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
43. Saini, N. et al. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. *PLoS Genet.* **12**, e1006385 (2016).
44. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
45. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072 (2015).
46. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
47. Kasar, S. & Brown, J. R. Mutational landscape and underlying mutational processes in chronic lymphocytic leukemia. *Mol. Cell. Oncol.* **3**, e1157667 (2016).
48. Matsuda, T., Kawanishi, M., Yagi, T., Matsui, S. & Takebe, H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res.* **26**, 1769–1774 (1998).
49. Brash, D. E. UV signature mutations. *Photochem. Photobiol.* **91**, 15–26 (2015).
50. Hill, K. A., Wang, J., Farwell, K. D. & Sommer, S. S. Spontaneous tandem-base mutations (TBM) show dramatic tissue, age, pattern and spectrum specificity. *Mutat. Res.* **534**, 173–186 (2003).
51. Garaycoechea, J. I. et al. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553**, 171–177 (2018).
52. Pfeifer, G. P. Formation and processing of UV photoproducts: effects of DNA sequence and chromatin environment. *Photochem. Photobiol.* **65**, 270–283 (1997).
53. Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol.* **26**, 52–64 (2016).
54. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
55. Huang, M. N. et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* **27**, 1475–1486 (2017).
56. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).
57. Olivier, M. et al. Modelling mutational landscapes of human cancers *in vitro*. *Sci. Rep.* **4**, 4482 (2014).
58. Szikriszt, B. et al. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* **17**, 99 (2016).
59. Zhivagui, M. et al. Experimental and pan-cancer genome analyses reveal widespread contribution of acrylamide exposure to carcinogenesis in humans. *Genome Res.* **29**, 521–531 (2019).
60. Zámbrorsky, J. et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**, 746–755 (2017).
61. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
62. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

#### PCAWG Mutational Signatures Working Group

Ludmil B. Alexandrov<sup>1</sup>, Erik N. Bergstrom<sup>1</sup>, Arnold Boot<sup>4,5</sup>, Paul Boutros<sup>27,28,29,30</sup>, Kin Chan<sup>31</sup>, Kyle R. Covington<sup>32</sup>, Akihiro Fujimoto<sup>32</sup>, Gad Getz<sup>2,3,21,22</sup>, Dmitry A. Gordenin<sup>8</sup>, Nicholas J. Haradhvala<sup>2,3</sup>, Mi Ni Huang<sup>4,5</sup>, S. M. Ashiqul Islam<sup>1</sup>, Marat Kazanov<sup>33,34,35</sup>, Jaegil Kim<sup>2</sup>, Leszek J. Klimczak<sup>32</sup>, Nuria Lopez-Bigas<sup>9,10,11</sup>, Michael Lawrence<sup>2,36,37</sup>, Iñigo Martincorena<sup>13</sup>, John R. McPherson<sup>4,5</sup>, Sandro Morganello<sup>13</sup>, Ville Mustonen<sup>17,18,19</sup>, Hideaki Nakagawa<sup>32</sup>, Alvin Wei Tian Ng<sup>4,5</sup>, Paz Polak<sup>2,3,22</sup>, Stephenie Prokopec<sup>29</sup>, Steven A. Roberts<sup>38,39</sup>, Steven G. Rozen<sup>4,5,23</sup>, Radhakrishnan Sabarinathan<sup>10,14,15</sup>, Natalie Saini<sup>8</sup>, Tatsuhiro Shibata<sup>40,41</sup>, Yuichi Shiraishi<sup>42</sup>, Michael R. Stratton<sup>13</sup>, Bin Tan Teh<sup>4,23,43,44,45</sup>, Ignacio Vázquez-García<sup>13,46,47,48</sup>, David A. Wheeler<sup>6,16</sup>, Yang Wu<sup>4,5</sup>, Fouad Yousif<sup>2</sup> & Willie Yu<sup>4,5</sup>

<sup>27</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.

<sup>28</sup>University of California Los Angeles, Los Angeles, CA, USA. <sup>29</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>30</sup>Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. <sup>31</sup>Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. <sup>32</sup>RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>33</sup>A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia. <sup>34</sup>Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia. <sup>35</sup>Skolkovo Institute of Science and Technology, Moscow, Russia. <sup>36</sup>Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>37</sup>Massachusetts General Hospital, Boston, MA, USA. <sup>38</sup>School of Molecular Biosciences, Washington State University, Pullman, WA, USA. <sup>39</sup>Center for Reproductive Biology, Washington State University, Pullman, WA, USA. <sup>40</sup>Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>41</sup>Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. <sup>42</sup>The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>43</sup>Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. <sup>44</sup>Institute of Molecular and Cell Biology, Singapore, Singapore. <sup>45</sup>Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Center Singapore, Singapore, Singapore. <sup>46</sup>Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>47</sup>Department of Statistics, Columbia University, New York, NY, USA. <sup>48</sup>Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK.



No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

These online methods contain an abridged description of the methodology used in the current manuscript; extensive details about the methodology we used are provided in Supplementary Note 2. Importantly, two independently developed computational frameworks (SigProfiler and SignatureAnalyzer) based on NMF were applied separately to the examined sets of mutational catalogues. SigProfiler and SignatureAnalyzer take different approaches for deciphering mutational signatures and for assigning each signature to each sample. By using two methods, we aimed to provide a perspective on the effect that different methodologies can have on the numbers of signatures generated, signature profiles and attributions. In addition to applying SigProfiler and SignatureAnalyzer to cancer data, the tools were also applied to realistic synthetic data with known solutions.

### Analysis of mutational signatures with SigProfiler

SigProfiler incorporates two distinct steps for identification of mutational signatures, based on the previously described methodology<sup>6,11,17</sup> (Extended Data Fig. 8). The first step (SigProfilerExtraction) encompasses a hierarchical de novo extraction of mutational signatures based on somatic mutations and their immediate sequence context, and the second step (SigProfilerAttribution) focuses on accurately estimating the number of somatic mutations associated with each extracted mutational signature in each sample. SigProfilerExtraction is an extension of a previous framework for the analysis of mutational signatures<sup>11,17</sup>. In brief, for a given set of mutational catalogues, the algorithm deciphers a minimal set of mutational signatures that optimally explains the proportion of each mutation type and estimates the contribution of each signature to each sample. More specifically, for each NMF iteration, SigProfilerExtraction minimizes a generalized Kullback–Leibler divergence constrained for nonnegativity (Supplementary Note 2). The algorithm uses multiple NMF iterations (in most cases 1,024) to identify the matrix of mutational signatures and the matrix of the activities of these signatures, as previously described<sup>17</sup>. The unknown number of signatures is determined by human assessment of the stability and accuracy of solutions for a range of values, as previously described<sup>17</sup>. The framework is applied hierarchically to increase its ability to find mutational signatures that generate few mutations or are present in few samples.

After signatures are discovered by SigProfilerExtraction, SigProfilerAttribution estimates their contributions to individual samples. For each examined sample, the estimation algorithm involves finding the minimum of the Frobenius norm of a constrained function using a nonlinear convex optimization programming solver using the interior-point algorithm<sup>63</sup>. See Supplementary Note 2 and Extended Data Fig. 8b for further details.

### Analysis of mutational signatures with SignatureAnalyzer

SignatureAnalyzer uses a Bayesian variant of NMF that infers the number of signatures through the automatic relevance determination technique and delivers highly interpretable and sparse representations for both signature profiles and attributions that strike a balance between data fitting and model complexity. Further details of the actual implementation of the computational approach have previously been published<sup>9,27,64</sup>. SignatureAnalyzer was applied by using a two-step signature extraction strategy using 1,536 pentanucleotide contexts for SBSs, 83 indel features and 78 DBS features. In addition to the separate extraction of SBS, indel and DBS signatures, we performed a ‘COMPOSITE’ signature extraction based on all 1,697 features (1,536 SBS + 78 DBS + 83 indel). For SBSs, the 1,536 SBS COMPOSITE signatures are preferred; for DBSs and indels, the separately extracted signatures are preferred.

In step 1 of the two-step extraction process, global signature extraction was performed for the samples with a low mutation burden ( $n = 2,624$ ). These excluded hypermutated tumours: those with putative polymerase epsilon (POLE) defects or mismatch repair defects (microsatellite instable tumours), skin tumours (which had intense UV-light mutagenesis) and one tumour with temozolomide (TMZ) exposure. Because the underlying algorithm of SignatureAnalyzer performs a stochastic search, different runs can produce different results. In step 1, we ran SignatureAnalyzer 10 times and selected the solution with the highest posterior probability. In step 2, additional signatures unique to hypermutated samples were extracted (again selecting the highest posterior probability over ten runs) while allowing all signatures found in the samples with low mutation burden, to explain some of the spectra of hypermutated samples. This approach was designed to minimize a well-known ‘signature bleeding’ effect or a bias of hyper- or ultramutated samples on the signature extraction. In addition, this approach provided information about which signatures are unique to the hypermutated samples, which was later used when attributing signatures to samples.

A similar strategy was used for signature attribution: we performed a separate attribution process for low- and hypermutated samples in all COMPOSITE, SBS, DBS and indel signatures. For downstream analyses, we preferred to use the COMPOSITE attributions for SBSs and the separately calculated attributions for DBSs and indels. Signature attribution in samples with a low mutation burden was performed separately in each tumour type (for example, Biliary–AdenoCA, Bladder–TCC, Bone–Osteosarc, and so on). Attribution was also performed separately in the combined microsatellite instable tumours ( $n = 39$ ), POLE ( $n = 9$ ), skin melanoma ( $n = 107$ ) and TMZ-exposed samples (syn11738314). In both groups, signature availability (which signatures were active, or not) was primarily inferred through the automatic relevance determination process applied to the activity matrix  $H$  only, while fixing the signature matrix  $W$ . The attribution in samples with a low mutation burden was performed using only signatures found in the step 1 of the signature extraction. Two additional rules were applied in SBS signature attribution to enforce biological plausibility and minimize a signature bleeding: (i) allow SBS4 (smoking signature) only in lung, head and neck cases; and (ii) allow SBS11 (TMZ signature) in a single GBM sample. This was enforced by introducing a binary, signature-by-sample signature indicator matrix  $Z$  (1, allowed; 0, not allowed), which was multiplied by the  $H$  matrix in every multiplication update of  $H$ . No additional rules were applied to indel or DBS signature attributions, except that signatures found in hypermutated samples were not allowed in samples with a low mutation burden.

### Application of SigProfiler and SignatureAnalyzer to synthetic data

Our goal was to evaluate SignatureAnalyzer and SigProfiler on realistic synthetic data to identify any potential limitations of these two methods. SignatureAnalyzer and SigProfiler were tested on 11 sets of synthetic data, encompassing a total of 64,400 synthetic samples, in which known signature profiles were used to generate catalogues of synthetic mutational spectra. We operationally defined ‘realistic’ data as those based on the characteristics of either SignatureAnalyzer’s or SigProfiler’s analysis of the PCAWG genome data. SignatureAnalyzer’s reference signature profiles were based on COMPOSITE signatures, consisting of 1,536 types of strand-agnostic SBSs in pentanucleotide context, 78 types of DBSs and 83 types of small indels, for a total of 1,697 mutation types. SigProfiler’s reference analysis was based on strand-agnostic SBSs in the context of one 5’ and one 3’ base. For each test, we generated two sets of realistic data: SigProfiler-realistic (based on SigProfiler’s reference signatures and attributions) and SignatureAnalyzer-realistic (based on SignatureAnalyzer’s reference signatures and attributions), as well as two other types of data that involved using SignatureAnalyzer profiles with SigProfiler attributions and vice versa.

A detailed description of each of the 11 sets of synthetic data and the results from applying SigProfiler and SignatureAnalyzer are provided in Supplementary Note 2.

### Analysis of clustered mutational signatures

Somatic SBSs were considered clustered if they had intermutational distances <1,000 bp. More specifically, for each sample, an SBS mutational catalogue was generated for substitutions that were <1,000 bp from another substitution. Subsequently, the set of SBS mutational catalogues containing clustered mutations underwent de novo extraction of mutational signatures. Any novel mutational signature (one that was not previously observed in the complete SBS catalogues) was reported as a clustered mutational signature.

### Better separation compared to COSMIC v.2 signatures

As described in the manuscript, all mutational signatures previously reported in COSMIC v.2 were confirmed in the new set of analyses with median cosine similarity of 0.95. However, the separation between the COSMIC v.2 mutational signatures ([https://cancer.sanger.ac.uk/cosmic/signatures\\_v2](https://cancer.sanger.ac.uk/cosmic/signatures_v2)) is much worse than the separation between the mutational signatures reported here. For example, in COSMIC v.2, signatures 5 and 16 had a cosine similarity of 0.90, making them hard to distinguish from one another. By contrast, in the current analysis, SBS5 and SBS16 have a cosine similarity of 0.65. This allows us to unambiguously assign SBS5 and SBS16 to different samples. In the current analysis, the larger number of samples has allowed the reduction of bleeding between signatures and has given more unique and easily distinguishable signatures. One can evaluate the overall separation of a set of mutational signatures by examining the distribution of cosine similarities between the signatures in the set. The signatures in COSMIC v.2 had a median cosine similarity of 0.238. By contrast, the current signatures have a much lower median cosine similarity of 0.098. This twofold reduction in similarity is highly statistically significant ( $P$  value  $9.1 \times 10^{-25}$ ) and indicates a better separation between the signatures in the current analysis.

### Correlations of mutational signature activity with age

Before evaluating the association between age and the activity of a mutational signature, all outliers for both age and numbers of mutations attributed to a signature in a cancer type were removed from the data. An outlier was defined as any value outside three standard deviations from the mean value. A robust linear regression model that estimated the slope of the line and whether this slope was significantly different from zero ( $F$  test;  $P$  value < 0.05) was performed using the MATLAB function `robustfit` (<https://www.mathworks.com/help/stats/robustfit.html>) with default parameters. The  $P$  values from the  $F$  tests were corrected using the Benjamini–Hochberg procedure for false discovery rates. Results are available at syn12030687 and syn20317940.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC and TCGA PCAWG Consortium are described in ref. <sup>2</sup>, and are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access information that could potentially identify participants, such as germline alleles and

the underlying sequencing data, researchers will need to apply to the TCGA data access committee via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC data access compliance office (<http://icgc.org/daco>) for the ICGC portion of the dataset. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization. For each mutational signature as extracted by SigProfiler, there is a ‘vignette’ that consists of plots and a short textual description at COSMIC (available at <https://cancer.sanger.ac.uk/cosmic/signatures/>). Beyond the core sequence data generated by the ICGC and TCGA PCAWG Consortium, other derived datasets were generated by the research reported in this paper. These derived datasets are available at Synapse (<https://www.synapse.org/#!Synapse:syn11726601/wiki/513478>), and are denoted by accession numbers (synXXXXXXXXX). All these datasets are mirrored at [https://dcc.icgc.org/releases/PCAWG/mutational\\_signatures/](https://dcc.icgc.org/releases/PCAWG/mutational_signatures/) with full links, filenames, accession numbers and descriptions as detailed in Supplementary Table 1. These datasets include (1) CSV files comprising all catalogues of observed mutational spectra that were used as input to signature extraction (syn11801889), (2) CSV files and plots of signatures extracted by SigProfiler (syn11738306) and SignatureAnalyzer (syn11738307), (3) CSV files with estimates of the numbers of mutations generated by each signature in individual tumours (syn11804065), (4) estimates of the probability that each signature was responsible for each mutational type (for example, CTG>CAG) in individual tumours (syn11804068) and (5) synthetic test input data plus the results of tests of signature extraction (discovery) on the synthetic test data (syn18497223). All derived datasets are open access, and can be downloaded without registration or logging in.

### Code availability

SigProfiler is available both as a MATLAB framework and as a Python package. In both cases, SigProfiler is a fully functional, free and open-source tool distributed under the permissive 2-Clause BSD License. SigProfiler in MATLAB can be downloaded from: <https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler>. SigProfiler in Python can be downloaded from: <https://github.com/Alexandrov-Lab/SigProfilerExtractor>. SignatureAnalyzer code is available at <https://github.com/broadinstitute/getzlab-SignatureAnalyzer> (github.com). The code used to generate the synthetic data and summarize SignatureAnalyzer and SigProfiler results is open source and freely available as the SynSig package: <https://github.com/steverozen/SynSig/tree/v0.2.0> under the GNU General Public License v.3.0. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at <https://dockstore.org/search?search=pcawg> under the GNU General Public License v.3.0, which allows for reuse and distribution.

63. Byrd, R. H., Hribar, M. E. & Nocedal, J. An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optim.* **9**, 877–900 (1999).

64. Tan, V. Y. & Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the  $\beta$ -divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).

**Acknowledgements** The results here are based in part on data generated by the TCGA research network (<http://cancergenome.nih.gov/>) and the ICGC and TCGA PCAWG network. This work was supported by Wellcome grant reference 206194 (M.R.S.), Singapore National Medical Research Council grants NMRC/CIRG/1422/2015 and MOH-000032/MOH-CIRG18may-0004 and the Singapore Ministry of Health via the Duke-NUS Signature Research Programmes (M.N.H., A.W.T.N., Y.W., A.B. and S.G.R.), US National Institute of Health Intramural Research Program Project Z1AES103266 (D.A.G.), the European Research Council Consolidator Grant 682398 (N.L.-B.), US National Cancer Institute U24CA143843 (D.A.W.) and Cancer Research UK Grand Challenge Award C98/A24032 (E.N.B., S.M.A.I., L.B.A. and M.R.S.). G.G. and J.K. were partially supported by the National Cancer Institute grants U24CA210999 and U24CA143845. G.G. was partially supported by the Paul C. Zamecnik, MD, Chair in Oncology at the Massachusetts General Hospital Cancer Center. N.J.H. and G.G. were partially supported by G.G.’s funds at the Broad Institute and Massachusetts General Hospital. N.J.H. was partially funded by the Molecular Biophysics Training Grant NIH/ NIGMS T32 GM008313 (PI: Venkatesh N. Murthy). We acknowledge the contributions of the many clinical networks across the ICGC

# Article

and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects. The members of the PCAWG Consortium are listed in Supplementary Note 1.

**Author contributions** The ICGC and TCGA contributed collectively to this work under the guidance of PCAWG Steering and Executive Committees, and the Ethics and Legal Working Group. The International Cancer Genome Consortium and TCGA tumour specific providers provided tumour and matched non-tumour samples, and the PCAWG Technical Working Group, the PCAWG Quality Control Working Group and the PCAWG Novel Somatic Mutation Calling Methods Working Group provided standardized mutation calls for the 2,780 PCAWG whole genomes. G.G., S.G.R. and M.R.S. were project leaders; L.B.A., G.G., S.G.R. and M.R.S. obtained funding for this study; L.B.A., J.K., N.J.H., G.G., S.G.R. and M.R.S. designed this study; M.N.H., A.W.T.N., A.B., E.N.B., J.R.M. and S.G.R. collected and prepared data for analysis; L.B.A., J.K., E.N.B. and S.M.A.I. created mutational signature analysis software; L.B.A., J.K., N.J.H.,

A.W.T.N., A.B., K.R.C., D.A.G., N.L.-B., L.J.K., S.M., R.S., D.A.W., V.M., G.G., S.G.R. and M.R.S. analysed data and reviewed results; L.B.A., J.K., N.J.H., G.G., S.G.R. and M.R.S. wrote the paper; L.B.A., J.K., N.J.H., M.N.H. and A.W.T.N. created figures; and Y.W. and S.G.R. generated synthetic data and benchmarked signature analysis software.

**Competing interests** G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTest and POLYSOLVER. All the other authors have no competing interests.

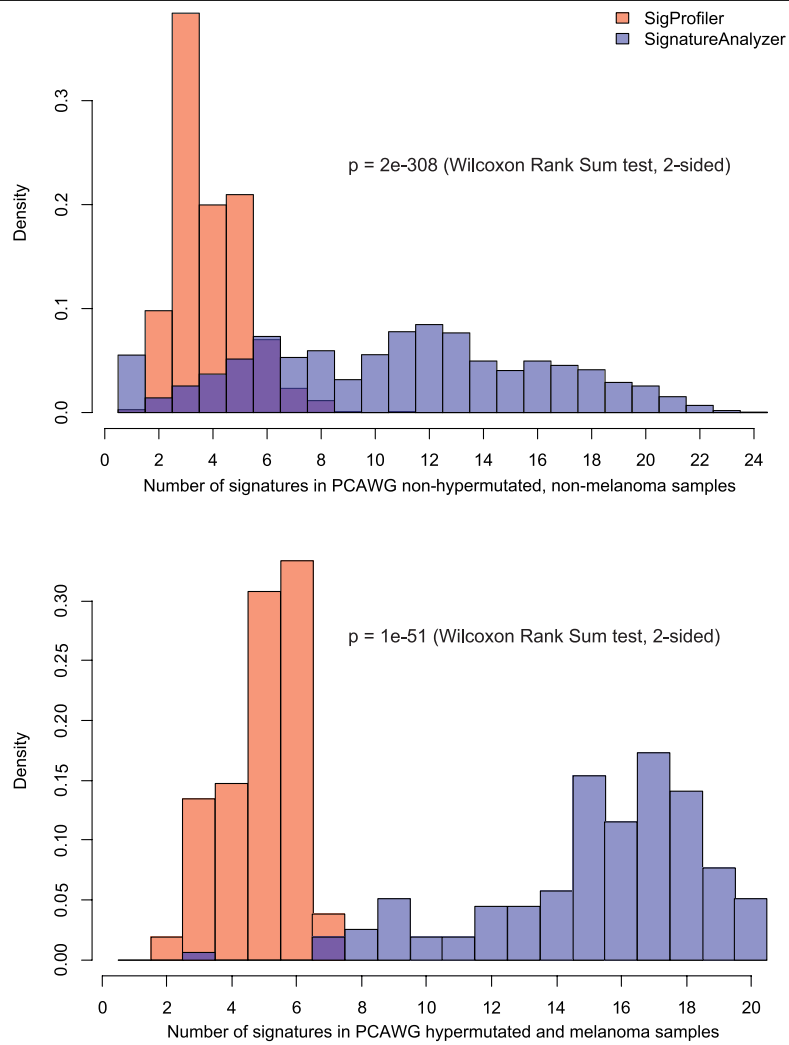
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1943-3>.

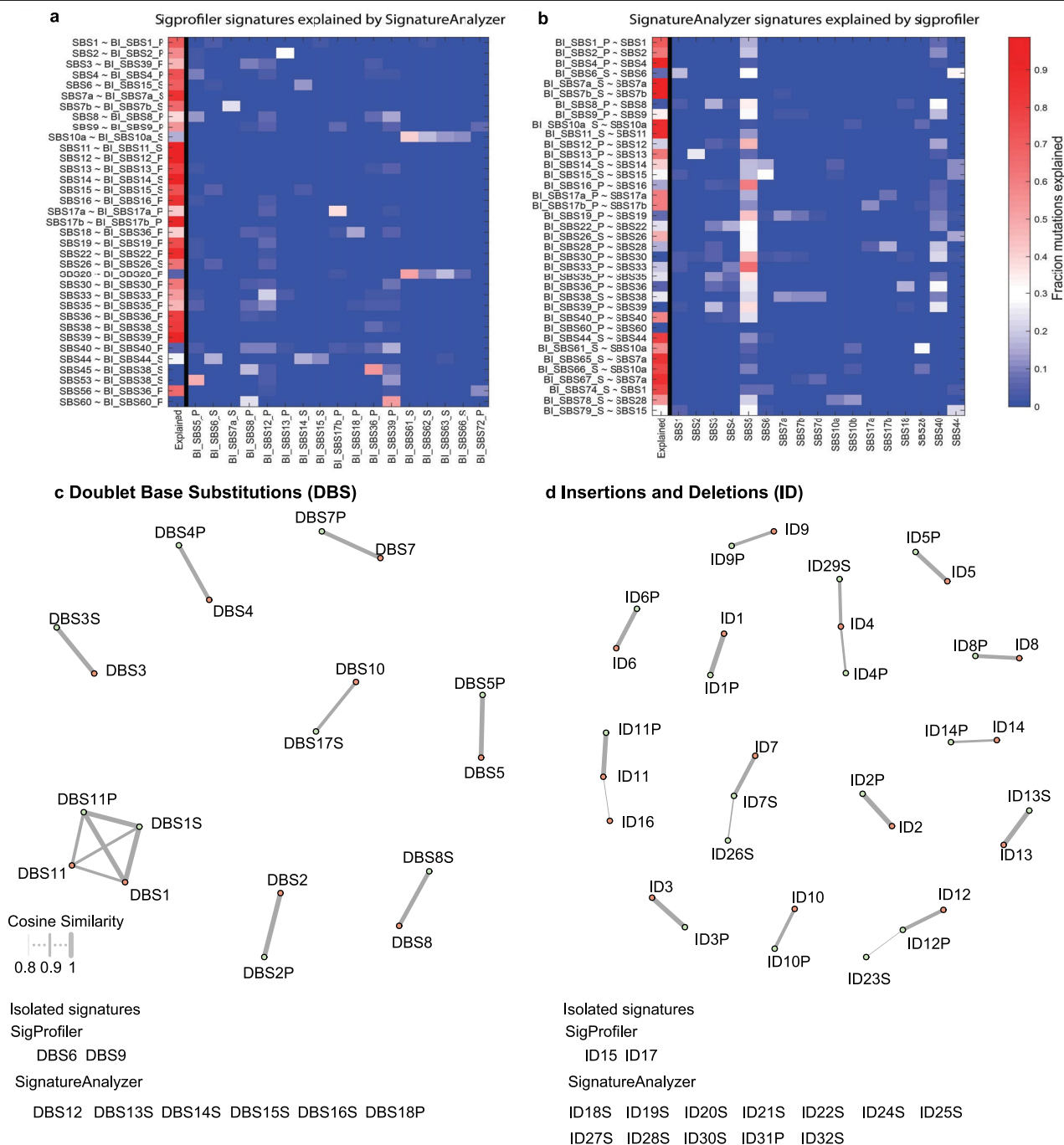
**Correspondence and requests for materials** should be addressed to S.G.R. or M.R.S.

**Peer review information** *Nature* thanks Arul Chinnaiyan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Histogram of the number of signatures attributed in each of 2,780 PCAWG samples by SigProfiler and SignatureAnalyzer. Hypermutated tumours and melanomas (156) are listed at syn11738314.**

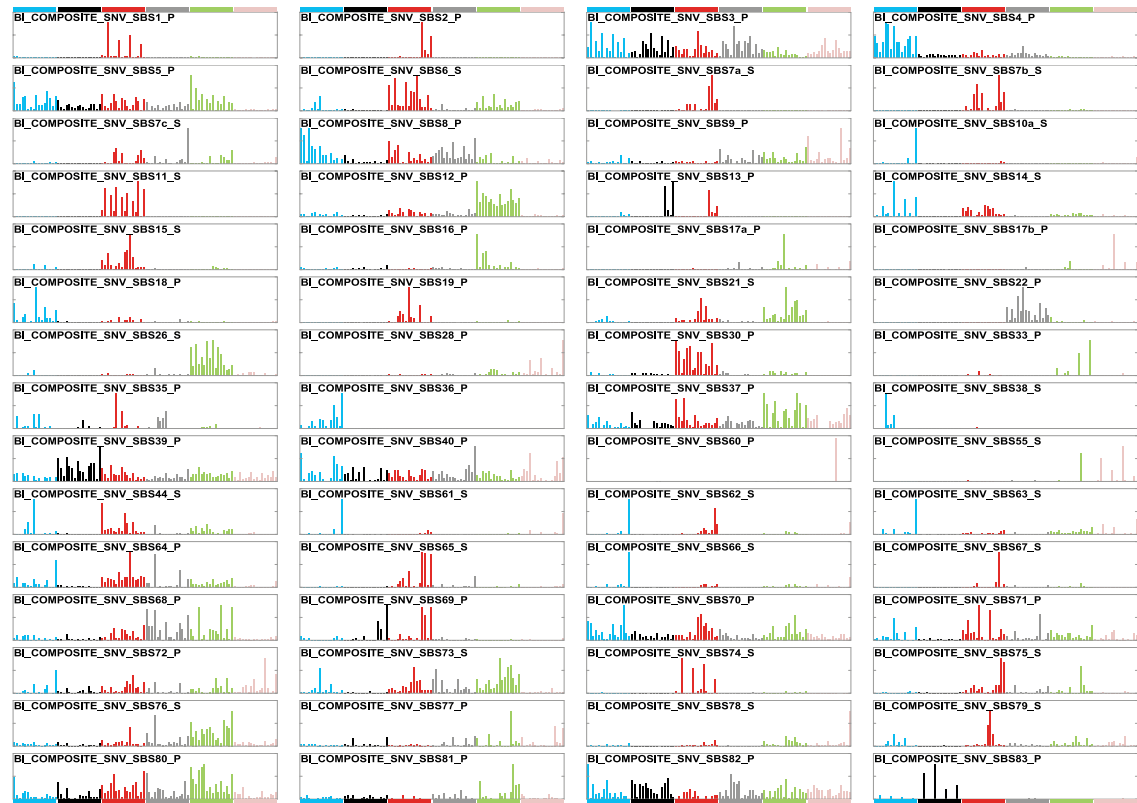


**Extended Data Fig. 2 | Comparisons between results of SigProfiler and SignatureAnalyzer. a, b.** Comparison of the attributions for corresponding SigProfiler (a) and SignatureAnalyzer (b) signatures. Each one of the SBS signatures extracted by SigProfiler and SignatureAnalyzer was paired with the signature of highest cosine similarity in the extraction by the other method (if one with  $>0.85$  cosine similarity exists). The first column of the plot corresponds to the fraction of mutations assigned by one method (summed across samples and mutation types) that was also assigned by the other method. The remaining mutations were then redistributed to the other signatures in the extraction, weighted by their relative probabilities of having been generated by each signature and the resulting fraction of mutations was then plotted. Signatures on the x axis are shown only if they contribute at least

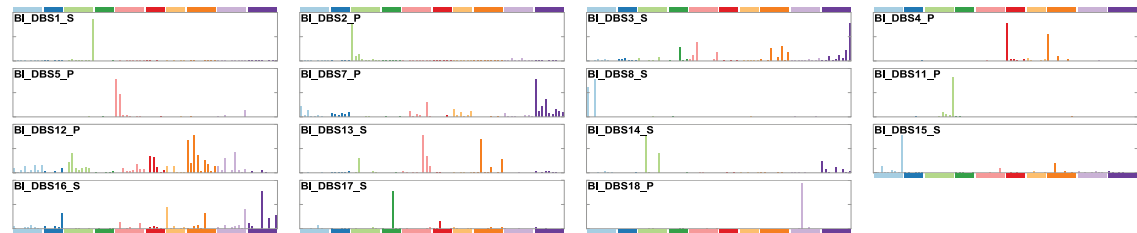
0.1 fraction of mutations to at least one signature on the y axis. **c, d.** Cosine similarities between SigProfiler and SignatureAnalyzer DBS (c) and indel (d) signatures. Brown nodes represent SigProfiler signatures; green nodes represent SignatureAnalyzer signatures. Matches with cosine similarities  $>0.8$  are shown as edges; the width of the edge indicates the strength of the similarity. The locations of the nodes have no meaning. Signatures with no matches of  $>0.8$  cosine similarity are shown below. SigProfiler ID15 and ID17 were extracted from data that were not analysed by SignatureAnalyzer. The suffix 'P' on a SignatureAnalyzer signature name indicates a signature extracted from non-hypermuted, non-melanoma tumours. The suffix 'S' on a SignatureAnalyzer signature name indicates a signature extracted from hypermutated or melanoma tumours.



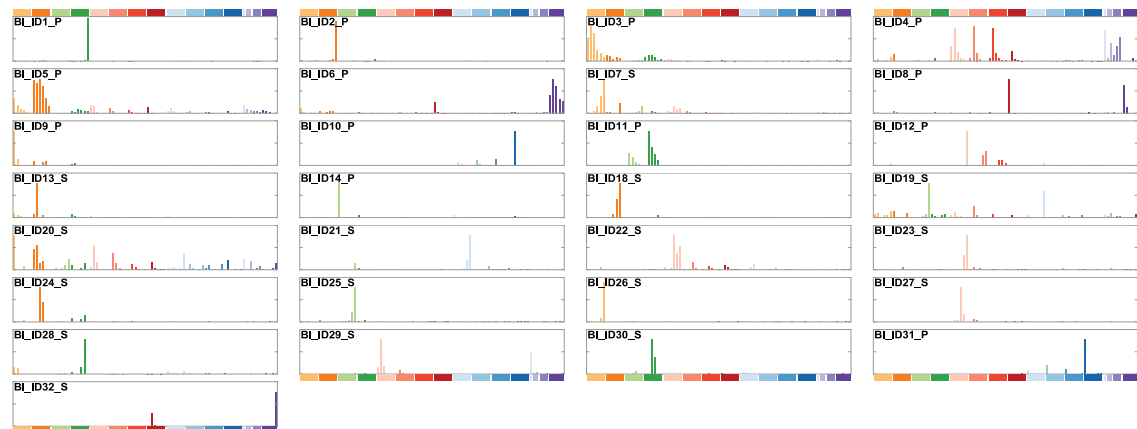
### SignatureAnalyzer reference SBS signatures



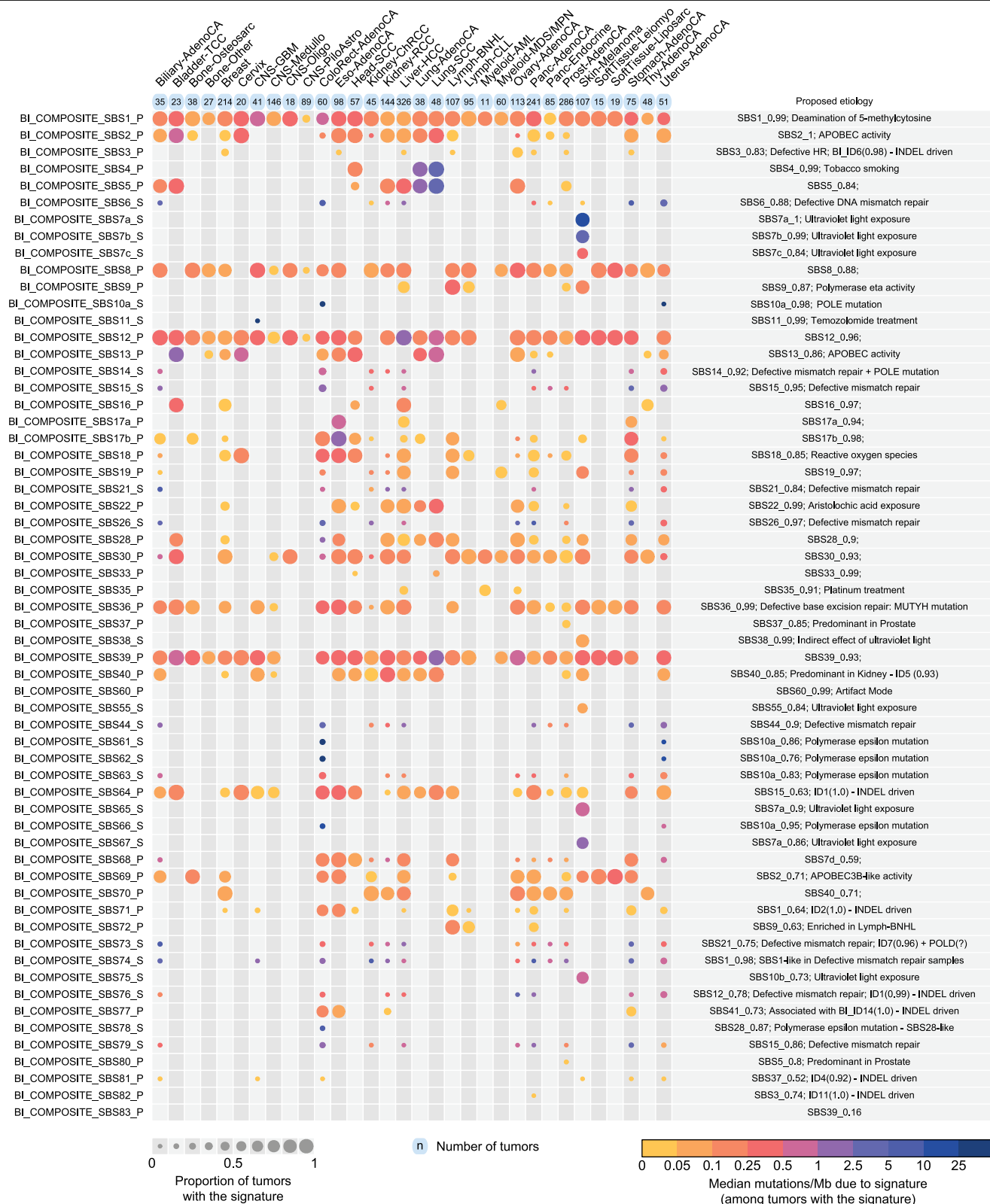
### SignatureAnalyzer reference DBS signatures



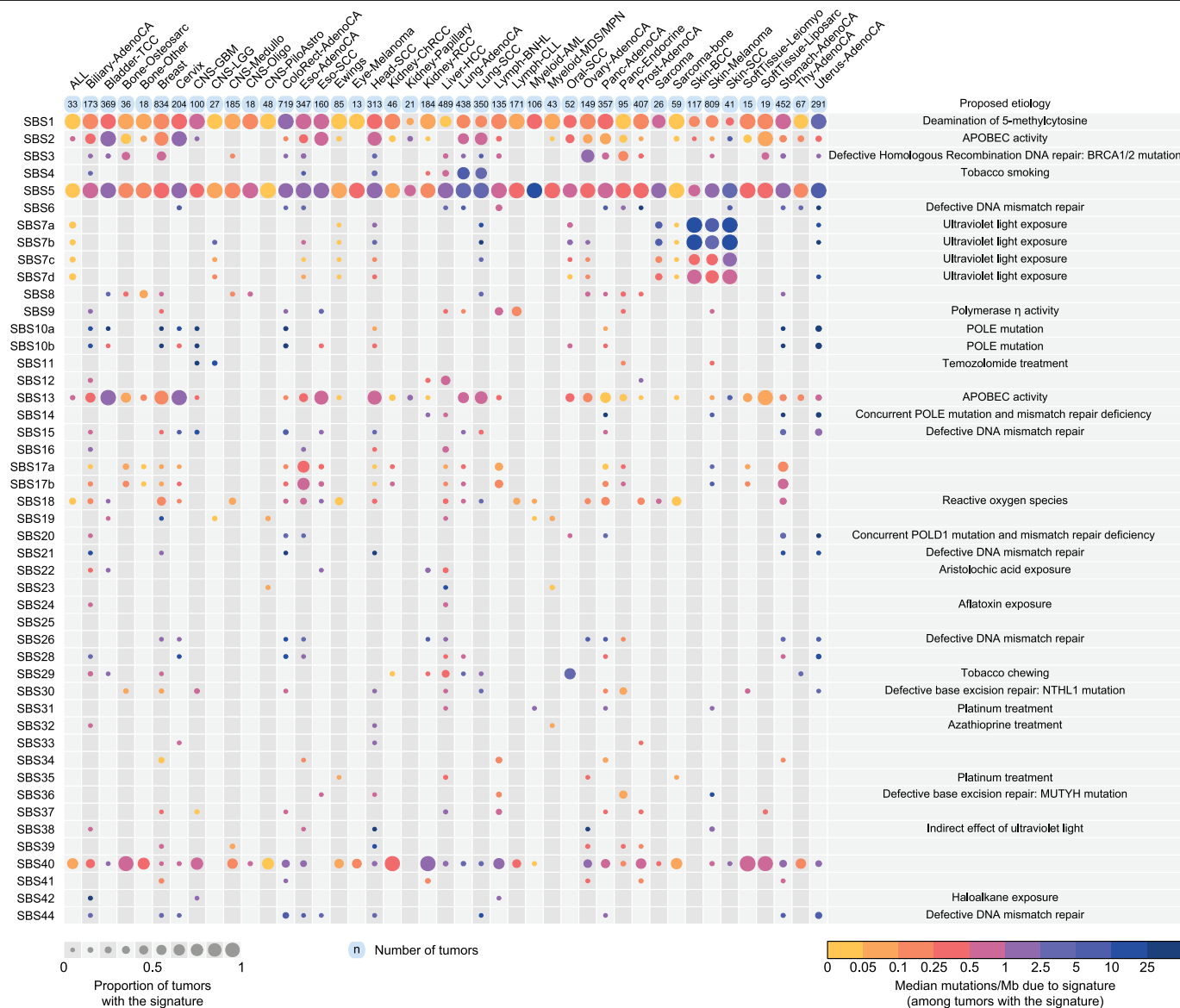
### SignatureAnalyzer reference ID signatures



**Extended Data Fig. 3 | SignatureAnalyzer reference signatures.** The classifications of each mutation type (SBS, 96 classes; DBS, 78 classes; and indels, 83 classes) are described in the main text.

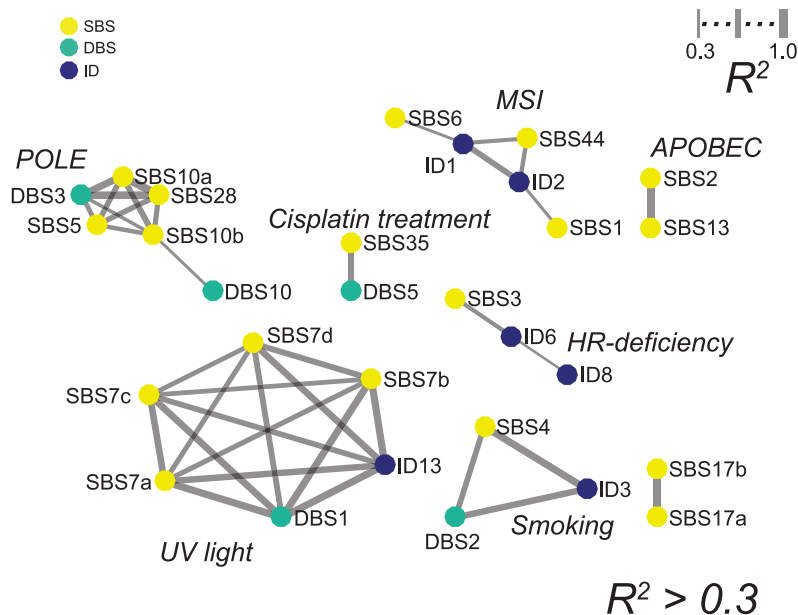


**Extended Data Fig. 4 | The number of SBS mutations attributed to each mutational signature for each cancer type over the PCAWG tumours by SignatureAnalyzer.** Conventions are as in Fig. 3; see this figure for explanation.

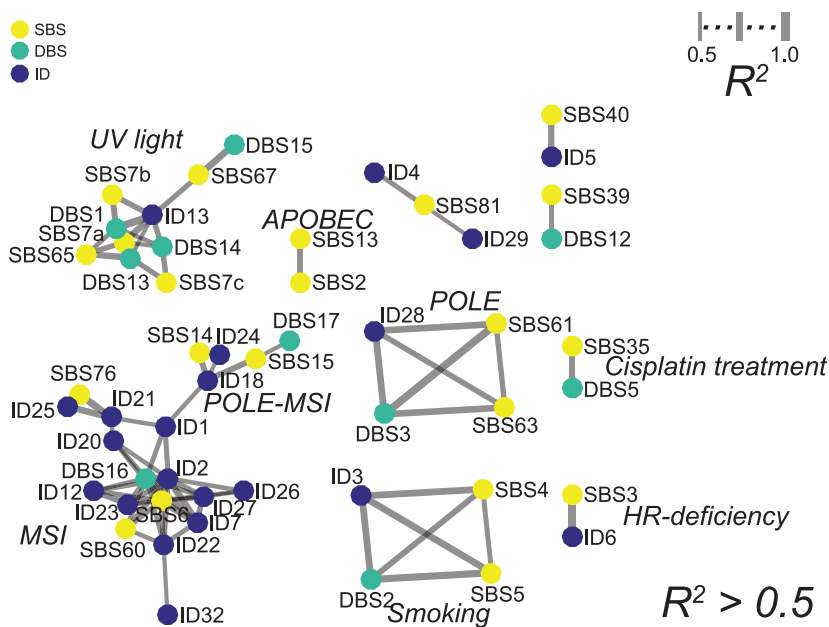


**Extended Data Fig. 5 | The number of SBS mutations attributed to each mutational signature to each cancer type over the complete set of PCAWG and non-PCAWG cancer samples analysed by SigProfiler.** Conventions are as in Fig. 3; see this figure for explanation.

## a SigProfiler



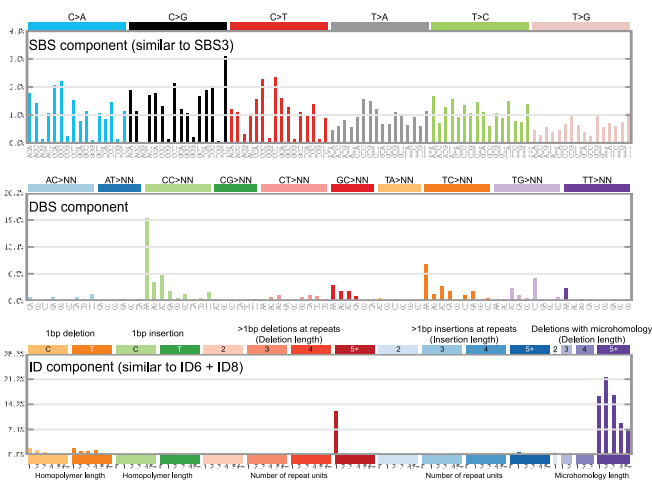
## b SignatureAnalyzer



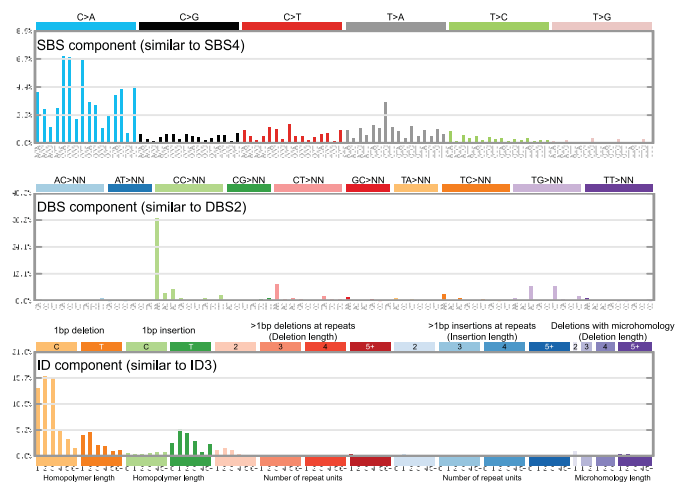
**Extended Data Fig. 6 | Associations between SBS, DBS and indel signature activities for SigProfiler and SignatureAnalyzer.** **a, b**, Each node represents an SBS (light green), DBS (dark green) or indel (black) signature. Any two signatures with sample attributions that significantly correlated with  $R^2 > 0.3$  (SigProfiler) (**a**) or  $> 0.5$  (SignatureAnalyzer) (**b**) are connected by edges. Edge

widths are proportional to the strength of the correlation. Signatures with no significant correlation to any other signature above the relevant threshold are not shown. Signature locations are fit for display purposes only, and do not indicate similarity.

### Composite-3



### Composite-4



### Composite-7a



### Composite-7b

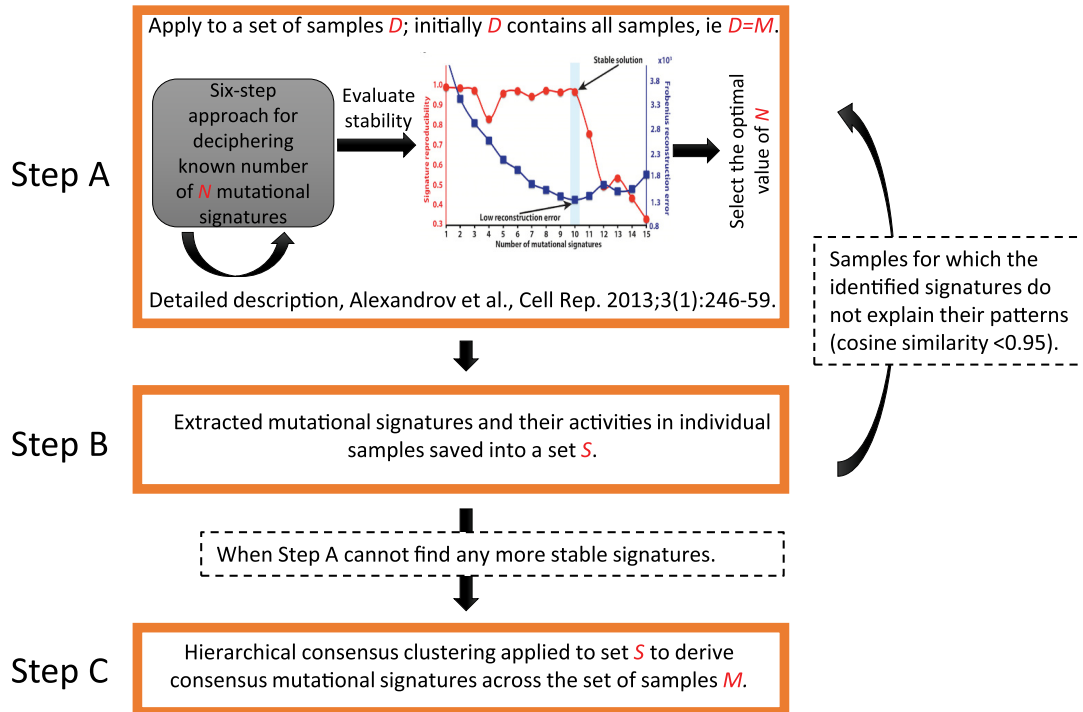


**Extended Data Fig. 7 | Mutational signatures extracted from the COMPOSITE feature set consisting of the concatenation of SBSs in pentanucleotide context, DBSs and indels.** For each of the 4 COMPOSITE mutational signatures shown, the top panel shows the SBS signature in pentanucleotide context (1,536 mutation classes) after being collapsed to

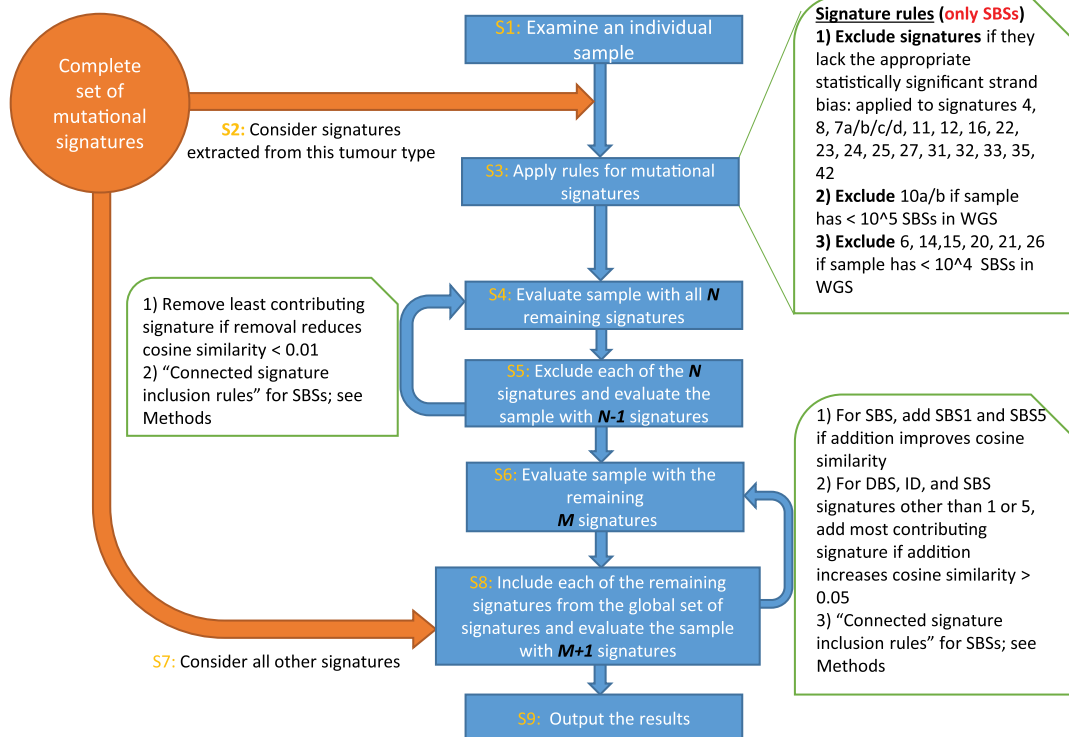
96 SBS mutation classes, the middle panel is the co-extracted DBS signature and the bottom panel is the co-extracted indel signature. There are similarities between the DBS portion of Composite-4 and DBS2, and between the indel portion of Composite-4 and ID3; other similarities are noted in the figure.



## a Extraction of mutational signatures



## b Attribution of activities of mutational signatures in samples



## Extended Data Fig. 8 | SigProfiler signature extraction and attribution.

A full description is provided in Supplementary Note 2. **a**, Procedure for extracting (discovering) mutational signatures. Step A, apply the approach to a set of samples  $D$ ; initially  $D$  contains all samples (that is,  $D=M$ ). This step has previously been described in detail<sup>17</sup>. Step B, solution evaluation and re-iteration. Extracted mutational signatures and their activities in individual samples are saved into a set ( $S$ ). The activity of any signature that does not increase the cosine similarity of a sample by > 0.01 was removed from the

sample (assigned a value of 0). Step A is repeated for all samples for which the identified signatures do not explain their patterns (cosine similarity < 0.95). The algorithm continues to step C when step A cannot find any stable signatures. Step C, clustering of mutational signatures. Hierarchical consensus clustering was applied to the set  $S$  to derive the consensus mutational signatures across the set of samples  $M$ . **b**, Attribution of activities of mutational signatures in samples.

**Extended Data Table 1 | The number of DBSs is proportional to the number of SBSs, with few exceptions**

Covariate (including cancer type)	Coefficient estimate	Coefficient std. error	t value	Pr(>  t )	Tumour count
(Intercept)	5.60E+00	8.80E+01	0.1	0.9	NA
SBS.count	3.70E-03	1.30E-04	29.8	<2e-16	NA
Biliary-AdenoCA	(reference)				35
Bladder-TCC	1.30E+01	1.40E+02	0.1	0.9	23
Bone-Benign	-6.30E+00	1.60E+02	0	1	16
Bone-Epith	2.20E+00	1.80E+02	0	1	11
Bone-Osteosarc	2.20E+00	1.20E+02	0	1	38
Breast-AdenoCA	6.20E+00	9.50E+01	0.1	0.9	198
Breast-DCIS	4.20E+00	3.10E+02	0	1	3
Breast-LobularCA	-8.20E+00	1.70E+02	0	1	13
Cervix-AdenoCA	-7.90E+00	3.80E+02	0	1	2
Cervix-SCC	-1.10E+01	1.50E+02	-0.1	0.9	18
CNS-GBM	-2.80E+01	1.20E+02	-0.2	0.8	41
CNS-Medullo	-7.00E+00	9.80E+01	-0.1	0.9	146
CNS-Oligo	-1.00E+01	1.50E+02	-0.1	0.9	18
CNS-PiloAstro	-5.90E+00	1.00E+02	-0.1	1	89
ColoRect-AdenoCA	-4.10E+02	1.10E+02	-3.7	3.00E-04	60
Eso-AdenoCA	-1.60E+01	1.00E+02	-0.2	0.9	98
Head-SCC	5.30E+01	1.10E+02	0.5	0.6	57
Kidney-ChRCC	-3.10E+00	1.20E+02	0	1	45
Kidney-RCC	5.60E+01	9.80E+01	0.6	0.6	144
Liver-HCC	7.80E+01	9.20E+01	0.8	0.4	326
Lung-AdenoCA	5.00E+02	1.20E+02	4.1	4.00E-05	38
Lung-SCC	5.80E+02	1.20E+02	5.1	4.00E-07	48
Lymph-BNHL	1.00E+01	1.00E+02	0.1	0.9	107
Lymph-CLL	-4.30E+00	1.00E+02	0	1	95
Myeloid-AML	-1.90E+00	1.80E+02	0	1	11
Myeloid-MDS	-8.00E+00	2.70E+02	0	1	4
Myeloid-MPN	-7.40E+00	1.10E+02	-0.1	0.9	56
Ovary-AdenoCA	3.60E+01	1.00E+02	0.4	0.7	113
Panc-AdenoCA	-8.30E-01	9.40E+01	0	1	241
Panc-Endocrine	-5.70E+00	1.00E+02	-0.1	1	85
Prost-AdenoCA	2.50E+00	9.30E+01	0	1	286
Skin-Melanoma	1.70E+03	1.00E+02	16.5	<2e-16	107
SoftTissue-Leiomyo	6.00E+00	1.60E+02	0	1	15
SoftTissue-Liposarc	7.80E+00	1.50E+02	0.1	1	19
Stomach-AdenoCA	-3.00E+01	1.10E+02	-0.3	0.8	75
Thy-AdenoCA	-4.80E+00	1.20E+02	0	1	48
Uterus-AdenoCA	-1.20E+02	1.10E+02	-1.1	0.3	51

The exceptions are colorectal adenocarcinoma (Colorect-AdenoCA), lung adenocarcinoma (Lung-AdenoCA), lung squamous cell carcinoma (Lung-SCC) and skin-melanoma, as analysed by the following linear regression (computed by an R function call): `glm(DBS.count ~ SBS.count + Cancer.Type)`. This function call fits a model in which the number of DBSs depends linearly on the number of SBSs and on the cancer type. *P* values associated with the coefficients are two-sided.

Extended Data Table 2 | Numbers of insertion and deletion mutations due to ID1, ID2 and all other indel signatures in hypermuted and non-hypermuted tumours

Signature	Hypermutators		Non-hypermutators		All Tumours	
	Count	Fraction	Count	Fraction	Count	Fraction
ID1	593,935	0.236	399,633	0.276	993,568	0.250
ID2	1,838,867	0.730	252,893	0.174	2,091,760	0.527
ID1+ID2	2,432,802	0.966	652,526	0.450	3,085,328	0.777
Other ID signatures	85,038	0.034	797,964	0.550	883,002	0.223
Total	2,517,840	1	1,450,490	1	3,968,330	1

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

## Software and code

Policy information about [availability of computer code](#)

### Data collection

The data in this study were those reported in <https://www.biorxiv.org/content/early/2017/07/12/162784.full.pdf+html> (the PCAWG marker paper) and in the publications cited at <https://www.synapse.org/#!Synapse:syn11801788>.

For the larger PCAWG Consortium, data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at <https://www.overture.bio>. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to [contact@overture.bio](mailto:contact@overture.bio).

### Data analysis

SigProfiler is available both as a MATLAB framework and as a Python package. In both cases, SigProfiler is fully functional, free, and open-source tool distributed under the permissive 2-Clause BSD License. SigProfiler in MATLAB can be downloaded from: <https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler> SigProfiler in Python can be downloaded from: <https://github.com/AlexandrovLab/SigProfilerExtractor>. SignatureAnalyzer code is available at <https://www.synapse.org/#!Synapse:syn11801492>. The code used to generate the synthetic data and summarize SignatureAnalyzer and SigProfiler results is open-source and freely available as the SynSig package: <https://github.com/stevenrozen/SynSig/tree/v0.2.0> under the GPLv3 license.

For the larger PCAWG Consortium, the workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: <https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files>. Individual software components are as follows: BWA-MEM v0.7.8-r455; DELLY v0.6.6; ACESeq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBbA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGRENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Derived data are available at <https://www.synapse.org/#!Synapse:syn11726601/wiki/513478>. All figures and extended data figures have associated raw data at that site.

For the larger PCAWG Consortium, WGS somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

From a statistical perspective this was an exploratory study, and there were no pre-defined hypothesis tests for which sample-size power calculations would have been appropriate. The sample size was determined by numbers of tumour genomes and exomes represented by publicly available somatic mutation data. These data consisted of the ICGC Pan Cancer whole genome mutation data, the TCGA MC3 whole exome mutation data, and additional mutation data as described in <https://www.synapse.org/#!Synapse:syn11801788>. This was an unsupervised analysis, and therefore we extracted as many signatures as possible from all the available data. This enabled a substantial increment over previously available sets of mutational signatures, especially with respect to double base substitution (DBS) signatures and insertion/deletion (ID) signatures.

For the larger PCAWG Consortium, the Consortium compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads.

We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014.

### Data exclusions

From a statistical perspective this was an exploratory study, and there were no pre-defined hypothesis tests for which pre-defined data exclusion criteria would have been appropriate. Therefore, no data were excluded from analysis by our algorithms.

For the larger PCAWG Consortium, after quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine).

### Replication

This was not an experimental study, and there were no experimental replicates.

For the larger PCAWG Consortium, in order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement.

### Randomization

There were no experimental groups in this study; the question of allocation to experimental groups is not applicable.

For the larger PCAWG Consortium, no randomisation was performed.



## Blinding

There was no allocation to experimental groups; the question of whether investigators were blinded to allocation is not applicable.

For larger PCAWG Consortium, no blinding was undertaken.

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

## Study description

Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).

## Research sample

State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.

## Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.

## Data collection

Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.

## Timing

Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.

## Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

## Non-participation

State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

## Randomization

If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

## Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

## Research sample

Describe the research sample (e.g. a group of tagged *Passer domesticus*, all *Stenocereus thurberi* within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

## Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

## Data collection

Describe the data collection procedure, including who recorded the data and how.

## Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

## Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

## Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

## Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

## Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? ☐ Yes ☐ No

## Field work, collection and transport

Field conditions	<i>Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).</i>
Location	<i>State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).</i>
Access and import/export	<i>Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).</i>
Disturbance	<i>Describe any disturbance caused by the study and how it was minimized.</i>

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	<i>Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Validation	<i>Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.</i>

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	<i>State the source of each cell line used.</i>
Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

## Palaeontology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	<i>For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.</i>
Wild animals	<i>Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<i>For the PCAWG Consortium data, patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-O-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced. The non-PCAWG analyses used previously published data.</i>
Recruitment	<i>For the PCAWG Consortium data, patients were recruited by the participating centres following local protocols.</i>
Ethics oversight	<i>For the PCAWG Consortium data, the Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>
Outcomes	<i>Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.</i>

## ChIP-seq

### Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links <i>May remain private before publication.</i>	<i>For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.</i>
Files in database submission	<i>Provide a list of all files available in the database submission.</i>
Genome browser session (e.g. <a href="#">UCSC</a> )	<i>Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.</i>

## Methodology

Replicates	<i>Describe the experimental replicates, specifying number, type and replicate agreement.</i>
Sequencing depth	<i>Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.</i>
Antibodies	<i>Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Peak calling parameters	<i>Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.</i>
Data quality	<i>Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.</i>
Software	<i>Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.</i>

## Flow Cytometry

### Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation	<i>Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.</i>
Instrument	<i>Identify the instrument used for data collection, specifying make and model number.</i>
Software	<i>Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.</i>
Cell population abundance	<i>Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.</i>
Gating strategy	<i>Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.</i>
<input type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

## Magnetic resonance imaging

### Experimental design

Design type	<i>Indicate task or resting state; event-related or block design.</i>
Design specifications	<i>Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.</i>
Behavioral performance measures	<i>State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).</i>

## Acquisition

Imaging type(s)	<i>Specify: functional, structural, diffusion, perfusion.</i>
Field strength	<i>Specify in Tesla</i>
Sequence & imaging parameters	<i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i>
Area of acquisition	<i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>
Diffusion MRI	<input type="checkbox"/> Used <input type="checkbox"/> Not used

## Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

## Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	<input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both
Statistic type for inference (See <a href="#">Eklund et al. 2016</a> )	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

## Models & analysis

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity
<input type="checkbox"/>	<input type="checkbox"/> Graph analysis
<input type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>
Multivariate modeling and predictive analysis	<i>Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.</i>



# Analyses of non-coding somatic drivers in 2,658 cancer whole genomes

<https://doi.org/10.1038/s41586-020-1965-x>

Received: 19 January 2018

Accepted: 2 December 2019

Published online: 5 February 2020

Open access

Esther Rheinbay<sup>1,2,3,73</sup>, Morten Muhlig Nielsen<sup>4,73</sup>, Federico Abascal<sup>5,73</sup>, Jeremiah A. Wala<sup>1,6,73</sup>, Ofer Shapira<sup>1,7,73</sup>, Grace Tiao<sup>1</sup>, Henrik Hornshøj<sup>4</sup>, Julian M. Hess<sup>1</sup>, Randi Istrup Juul<sup>4</sup>, Ziao Lin<sup>1,8</sup>, Lars Feuerbach<sup>9</sup>, Radhakrishnan Sabarinathan<sup>10,11</sup>, Tobias Madsen<sup>4</sup>, Jaegil Kim<sup>1</sup>, Loris Mularoni<sup>10,11</sup>, Shimin Shuai<sup>12,13</sup>, Andrés Lanzós<sup>14,15,16</sup>, Carl Herrmann<sup>17,18</sup>, Yosef E. Maruvka<sup>1,2</sup>, Ciyue Shen<sup>19,20</sup>, Samirkumar B. Amin<sup>21,22</sup>, Pratiti Bandopadhyay<sup>1,7</sup>, Johanna Bertl<sup>4</sup>, Keith A. Boroevich<sup>23</sup>, John Busanovich<sup>1,7</sup>, Joana Carlevaro-Fita<sup>14,15,16</sup>, Dimple Chakravarty<sup>24,25</sup>, Calvin Wing Yiu Chan<sup>17,26</sup>, David Craft<sup>27</sup>, Priyanka Dhingra<sup>28,29</sup>, Klev Diamanti<sup>30</sup>, Nuno A. Fonseca<sup>31</sup>, Abel Gonzalez-Perez<sup>10,11</sup>, Qianyun Guo<sup>32</sup>, Mark P. Hamilton<sup>33</sup>, Nicholas J. Haradvala<sup>1,2</sup>, Chen Hong<sup>9,26</sup>, Keren Isaev<sup>12,34</sup>, Todd A. Johnson<sup>23</sup>, Malene Juul<sup>4</sup>, Andre Kahles<sup>35</sup>, Abdullah Kahraman<sup>36</sup>, Youngwook Kim<sup>37</sup>, Jan Komorowski<sup>30,38</sup>, Kiran Kumar<sup>1,7</sup>, Sushant Kumar<sup>39</sup>, Donghoon Lee<sup>39</sup>, Kjong-Van Lehmann<sup>35</sup>, Yilong Li<sup>40,41</sup>, Eric Minwei Liu<sup>28,29</sup>, Lucas Lochovsky<sup>42</sup>, Keunchil Park<sup>37</sup>, Oriol Pich<sup>10,11</sup>, Nicola D. Roberts<sup>41</sup>, Gordon Saksena<sup>1</sup>, Steven E. Schumacher<sup>1,7</sup>, Nikos Sidiropoulos<sup>43</sup>, Lina Sieverling<sup>9,26</sup>, Nasa Sinnott-Armstrong<sup>44</sup>, Chip Stewart<sup>1</sup>, David Tamborero<sup>10,11</sup>, Jose M. C. Tubio<sup>45,46,47</sup>, Husen M. Umer<sup>30</sup>, Liis Uusküla-Reimand<sup>48,49</sup>, Claes Wadelius<sup>50</sup>, Lina Wadi<sup>12</sup>, Xiaotong Yao<sup>51</sup>, Cheng-Zhong Zhang<sup>52,53</sup>, Jing Zhang<sup>39</sup>, James E. Haber<sup>54</sup>, Asger Hobolth<sup>32</sup>, Marcin Imielinski<sup>51,55</sup>, Manolis Kellis<sup>1,56</sup>, Michael S. Lawrence<sup>1,2</sup>, Christian von Mering<sup>36</sup>, Hidewaki Nakagawa<sup>57</sup>, Benjamin J. Raphael<sup>58</sup>, Mark A. Rubin<sup>59,60,61</sup>, Chris Sander<sup>19,20</sup>, Lincoln D. Stein<sup>12,13</sup>, Joshua M. Stuart<sup>62</sup>, Tatsuhiko Tsunoda<sup>23,63,64</sup>, David A. Wheeler<sup>65</sup>, Rory Johnson<sup>14,16</sup>, Jüri Reimand<sup>12,34</sup>, Mark Gerstein<sup>39,42,66</sup>, Ekta Khurana<sup>28,29,60,61</sup>, Peter J. Campbell<sup>5,41</sup>, Núria López-Bigas<sup>10,11,67</sup>, PCAWG Drivers and Functional Interpretation Working Group<sup>68</sup>, PCAWG Structural Variation Working Group<sup>68</sup>, Joachim Weischenfeldt<sup>43,69,74\*</sup>, Rameen Beroukhi<sup>1,6,70,74\*</sup>, Iñigo Martincorena<sup>5,74\*</sup>, Jakob Skou Pedersen<sup>4,32,74\*</sup>, Gad Getz<sup>1,2,3,71,74\*</sup> & PCAWG Consortium<sup>72</sup>

The discovery of drivers of cancer has traditionally focused on protein-coding genes<sup>1–4</sup>. Here we present analyses of driver point mutations and structural variants in non-coding regions across 2,658 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium<sup>5</sup> of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). For point mutations, we developed a statistically rigorous strategy for combining significance levels from multiple methods of driver discovery that overcomes the limitations of individual methods. For structural variants, we present two methods of driver discovery, and identify regions that are significantly affected by recurrent breakpoints and recurrent somatic juxtapositions. Our analyses confirm previously reported drivers<sup>6,7</sup>, raise doubts about others and identify novel candidates, including point mutations in the 5' region of *TP53*, in the 3' untranslated regions of *NFKB1* and *TOB1*, focal deletions in *BRD4* and rearrangements in the loci of *AKR1C* genes. We show that although point mutations and structural variants that drive cancer are less frequent in non-coding genes and regulatory sequences than in protein-coding genes, additional examples of these drivers will be found as more cancer genomes become available.

Previous large-scale sequencing projects have identified many putative cancer genes, but most efforts have concentrated on mutations and copy-number alterations in protein-coding genes, mainly using whole-exome sequencing and single-nucleotide polymorphism arrays<sup>1–4</sup>. Whole-genome sequencing has made it possible to systematically survey non-coding regions for potential driver events, including

single-nucleotide variants (SNVs), small insertions and deletions (indels) and larger structural variants. Whole-genome sequencing enables the precise localization of structural variant breakpoints and connections between distinct genomic loci (juxtapositions). Although previous whole-genome sequencing analyses of modestly sized cohorts have revealed candidate non-coding regulatory driver events<sup>8–15</sup>,

The list of affiliations appears at the end of the paper.

the frequency and functional implications of these events remain understudied<sup>6,7,13,16,17</sup>.

Driver identification remains a far greater challenge in non-coding regions than in coding genes, owing to sequencing and mapping artefacts, poorly understood localized hypermutation processes<sup>14,18,19</sup>, incomplete annotation of regulatory regions, inaccurate estimation of the background mutation rate and the unknown functional effect of non-coding mutations. The discovery of drivers from structural variants is further complicated by their sparsity, the lack of obvious neutral events to build background models and their complex functional effects. Adequate statistical methods that address these issues are needed to reliably identify non-coding drivers.

The ICGC and TCGA PCAWG effort, which has collected and systematically analysed cancer genome sequences from 2,658 patients across 38 types of cancer<sup>5</sup>, offers an opportunity to characterize putative non-coding driver events that cannot be found using data from whole-exome sequencing or single-nucleotide polymorphism arrays. Here we describe a comprehensive search for non-coding somatic drivers. For point mutations (SNVs and indels), we combine results from multiple driver-discovery algorithms and, by carefully evaluating the significant hits, reveal that recurrent artefacts and poorly understood mutational processes have led to common false positives among previously reported non-coding drivers. For structural variants, we introduce two new methods for identifying both regions with significantly recurrent breakpoints (SRBs) and with significantly recurrent juxtapositions (SRJs), accounting for genomic heterogeneity in the rates of DNA break and repair and the three-dimensional architecture of the genome. Finally, to assess the potential for future non-coding driver discoveries, we quantify our statistical power in the PCAWG dataset and estimate the overall excess of point mutations in non-coding regulatory regions around known cancer genes.

### Hotspot mutations across cancer types

Many protein-coding driver mutations occur in single-site 'hotspots'. In the PCAWG dataset, only 12 single-nucleotide positions were mutated in >1%, and 106 in >0.5%, of patients (Extended Data Fig. 1a, Methods). Although protein-coding regions span only about 1% of the genome, 15 out of 50 (30%) of the most frequently mutated sites were well-studied hotspots in cancer genes (*KRAS*, *BRAF*, *PIK3CA*, *TP53* and *IDH1*) (Fig. 1a, Extended Data Fig. 1b), along with the two canonical *TERT* promoter hotspots<sup>6,7</sup>.

The remaining non-coding hotspots could be attributed to the following localized mutational processes associated with passenger events: (i) damage from ultraviolet (UV) light and impaired nucleotide excision repair in melanoma at sites occupied by transcription factors<sup>5,18–20</sup>; (ii) somatic hypermutation by activation-induced cytosine deaminase (AID) in B-cell non-Hodgkin lymphoma (Lymph–BNHL) and chronic lymphocytic leukaemia (Lymph–CLL); (iii) palindromic sequence contexts believed to form hairpin DNA structures targeted by APOBEC enzymes (in an intron of *GPR126* (also known as *ADGRG6*) and the *PLEKHS1* promoter)<sup>10</sup>; and (iv) presumed technical artefacts (Fig. 1a, Supplementary Note 1). These findings suggest that—besides *TERT* promoter events—non-coding single-site hotspot drivers are infrequent or fall in regions with low sensitivity to detect mutations.

### Discovery of point-mutation drivers

To identify recurrently mutated genomic elements, we first analysed somatic SNVs and indels in protein-coding regions, RNA genes (long and short non-coding RNAs and microRNAs (miRNAs)), and regulatory regions (promoters, 5' untranslated regions (UTRs), 3' UTRs and enhancers), totalling about 4% of the genome (Extended Data Fig. 2a–c, Methods, Supplementary Table 1). We analysed 2,583 tumours from 27 individual tumour types, and 15 meta-cohorts that grouped cancers

by tissue of origin or organ system (Extended Data Fig. 2d, Methods). We identified candidate drivers—that is, cohort–element combinations with  $Q < 0.1$  (10% false discovery rate (FDR))—by integrating 13 discovery algorithms, circumventing biases introduced by any one method (Extended Data Figs. 2e, 11, Supplementary Tables 2, 3, Supplementary Note 2). We benchmarked this approach by evaluating its ability to detect 603 known cancer genes (from the Cancer Gene Census (CGC)<sup>21</sup>, v.80), and found that combining methods improved performance compared to single algorithms (Extended Data Fig. 3a, b, Methods). Overall, we identified 1,294 significant hits that involved 520 unique candidates (Supplementary Tables 4, 5).

### Filtering the significant hits

Even after conservative FDR control, false-positive 'driver' loci can remain, owing to inaccurate background models, sequencing and mapping artefacts, or local increases in mutations due to unaccounted-for mutational processes. We therefore systematically filtered the candidate driver elements on the basis of technical and biological criteria, followed by careful review (Extended Data Fig. 3c, Methods, Supplementary Note 3). Examples of filtered elements include the promoters of *PIMI* (lymphoid tumours) and *RPL13A* (melanoma) because of associations with localized AID and UV-light mutational processes, respectively; *PLEKHS1*, *GPR126*, *TBC1D12* and *LEPROTL1* because of palindromic APOBEC target sequences<sup>9,10</sup>; and the *WDR74* 5' UTR and promoter<sup>8,10,14</sup>, owing to mapping problems detected in downstream manual review (Supplementary Table 5, Supplementary Note 4). In combination, filtering and reapplying FDR control discarded 589 out of 1,294 (46%) of the original cohort–element hits and 341 out of 520 (66%) unique elements (Extended Data Fig. 3c, Supplementary Tables 4, 5).

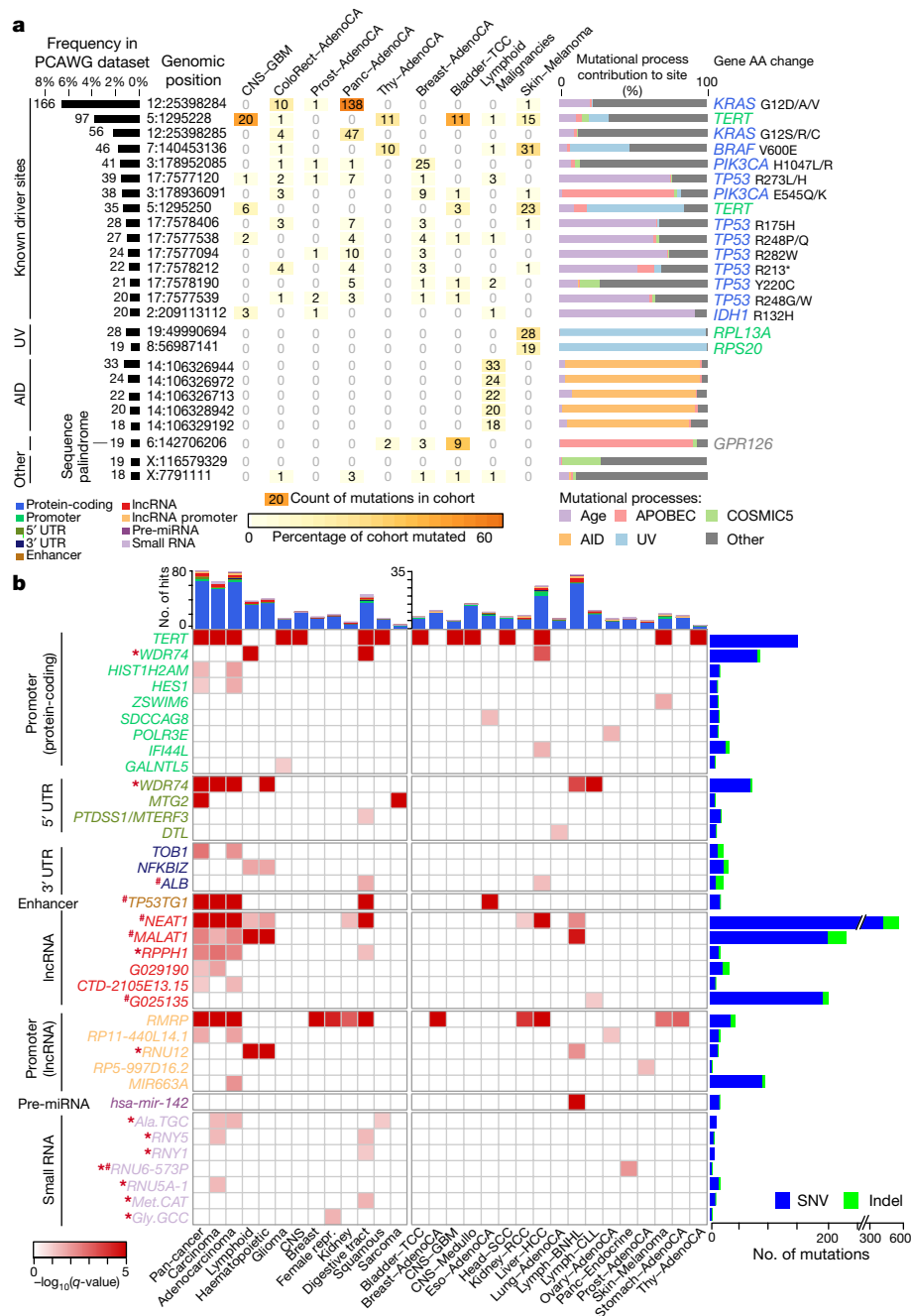
### Candidate coding and non-coding drivers

Our stringent combination and filtering strategy yielded 705 hits in 179 genomic elements: 602 hits in 143 protein-coding genes and 103 hits in non-coding elements. We observed wide variability across different types of cancer, from one hit in clear-cell renal cancer to 80 in the pan-cancer meta-cohort (Fig. 1b, Supplementary Tables 4, 5). Although most candidate drivers gained significance in larger meta-cohorts, some genes—such as *DAXX* (pancreatic endocrine tumour), *NRAS* (melanoma), *SPOP* (prostate adenocarcinoma), *FGFR1* (pilocytic astrocytoma) and *MIR142* (Lymph–BNHL)—scored higher in individual tumour types (Extended Data Fig. 3d). These results emphasize the trade-off between limiting driver discovery analyses to particular types of tumour and maximizing cohort size.

The candidate coding drivers we identified agreed with previous results: of the 143 genes that were significant in at least 1 cohort, 69% are in the CGC and nearly all have previously been implicated in cancer. In contrast to large whole-exome sequencing datasets, the fewer patients per cancer type in this dataset provided power sufficient only to detect genes with the strongest signal. We found 116 additional hits in 84 unique elements that were 'near significance' ( $0.1 < Q < 0.25$ ). Fifty-one per cent of the 63 unique protein-coding genes in this set are in the CGC, which suggests that they would have been discovered in larger cohorts (Supplementary Table 4).

To nominate a significant non-coding element as a candidate driver, we reviewed the supporting evidence from the mutation calls, additional genomic data (chromosomal breakpoints, copy number, loss-of-heterozygosity and expression data), cancer gene databases and the literature (Methods, Supplementary Tables 6, 10). We describe the key candidates below, and in Supplementary Note 4.

The *TERT* promoter was the most frequently mutated non-coding driver in this dataset (14 cohorts) (Fig. 1b), and these mutations were strongly associated with higher *TERT* expression, as has previously been reported<sup>9</sup> (Extended Data Fig. 4a, Supplementary Table 10).



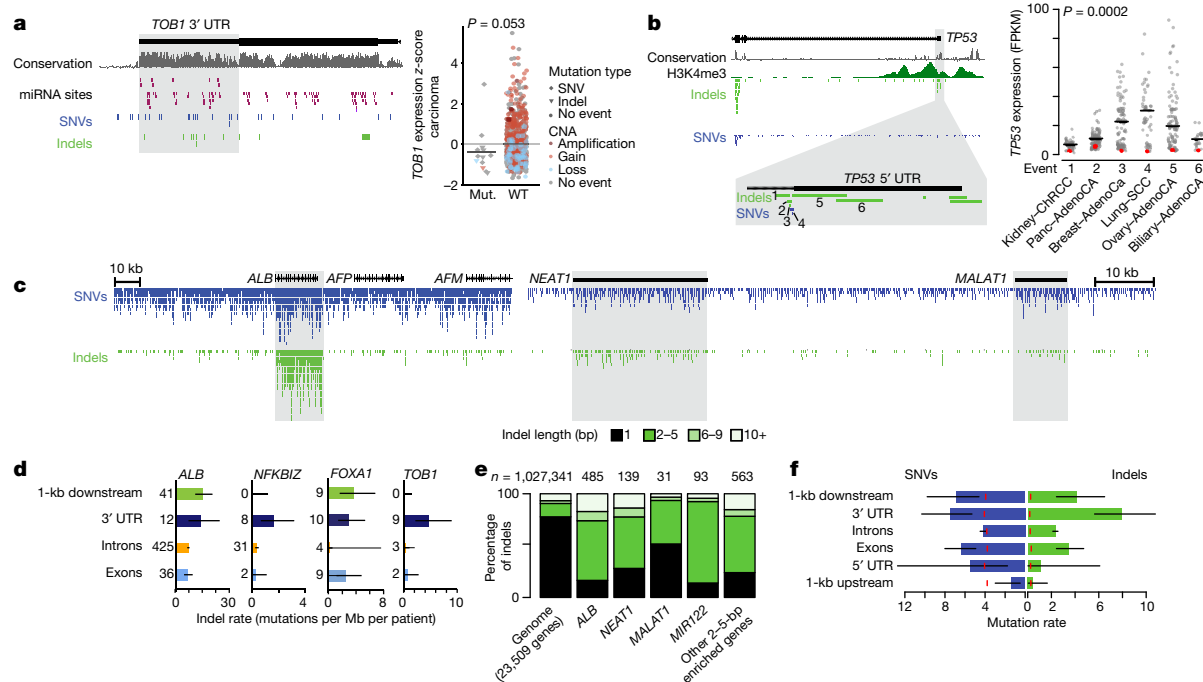
**Fig. 1 | Non-coding point mutations in PCAWG. a**, The bar chart (left) shows the total number of patients across PCAWG with mutations at a particular genomic hotspot (chromosome:position). The top 25 hotspots are grouped as known drivers or induced by mutational processes. The table (middle) shows the frequency of mutations across a subset of PCAWG cohorts. Lymphoid malignancies comprise Lymph–BNHL and Lymph–CLL. The stacked bar chart (right) shows the contribution of mutational processes to the hotspot mutations (Methods). Gene names are given when hotspots overlap functional elements (colour-coded), with amino acid (AA) alterations for protein-coding genes (solidus denotes substitution with any one of the indicated amino acids). Extended Data Fig. 1b shows the top 50 hotspots, and all cohorts. **b**, Significant non-coding elements ( $Q < 0.1$  of Brown's combined  $P$  values of up to 13 driver

discovery methods; Methods) identified before manual review in cohorts with at least one hit. Colour represents significance levels. Details are provided in Supplementary Table 5. \*Potential technical artefact; #targets affected by mutational processes. AdenoCA, adenocarcinoma; CNS, central nervous system; Eso, oesophageal; GBM, glioblastoma; HCC, hepatocellular carcinoma; Medullo, medulloblastoma; Panc, pancreatic; Prost, prostate; RCC, renal cell carcinoma; Repr., reproductive organs; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; Thy, thyroid. *HIST1H2AM* is also known as *H2AC17*; *Ala.TGC* as *TRA-TGC3-1*; *Met.CAT* as *TRM-CAT1-1*; and *Gly.GCC* as *TRG-GCC2-3*. *PTDSS1/MTERF3* denotes that 5' UTR mutations in *PTDSS1* also overlap the *MTERF3* promoter.

Mutations in the promoter and/or 5' UTR of *MTG2* (which encodes a GTPase involved in the mitochondrial ribosome) were associated with an expression of *MTG2* that was marginally significantly lower, in both the pan-cancer ( $P = 0.036$ , fold difference = 0.8) and carcinoma ( $P = 0.029$ , fold difference = 0.8) meta-cohorts (Extended Data Figs. 4a,

5a). Mutations in the 5' UTR have previously been shown to decrease *MTG2* expression in vitro<sup>22</sup>.

Recurrent somatic events were identified in the 3' UTRs of *TOB1* (carcinoma and pan-cancer meta-cohorts), *NFKBIZ* (lymphomas) and *ALB* (liver cancer) (Fig. 1b). *TOB1* encodes an anti-proliferation regulator



**Fig. 2 | Newly identified non-coding driver candidates and localized transcription-associated mutational process. a**, Recurrent mutations and associated gene expression in the highly conserved *TOB1* 3' UTR. Tracks showing conservation score (PhyloP, grey), miRNA-binding sites (TargetScan (top track) and Ago-Clip (bottom track)), and observed SNVs (blue) and indels (green). Expression of *TOB1* in mutated ( $n = 13$ ) and wild-type ( $n = 886$ ) cases (right).  $P$  value based on two-sided Wilcoxon rank-sum test. Bars represent means. CNA, copy-number alteration. **b**, Indels and SNVs overlapping the *TP53* 5' region and their effect on gene expression. H3K4me3 from the GM12878 cell line (ENCODE). Event numbers match with gene expression in the right panel (red dot, mutated sample; black bar, median).  $P$  value represents Fisher's combination of permutation tests within each tumour type. ChrRCC, chromophobe renal cell carcinoma; FPKM, fragments per kilobase of

transcript per million mapped reads. **c**, Overall pan-cancer distribution of indels and SNVs in *ALB*, *NEAT1* and *MALAT1* genomic loci (lymphoid tumour samples were excluded owing to AID). **d**, Quantification of average indel rates for genes with significantly mutated 3' UTRs. Error bars represent 95% binomial confidence intervals. **e**, Contribution of indels of different sizes in: all protein-coding and long non-coding RNA genes; *ALB*; *NEAT1*; *MALAT1*; *MIR122*; and the remaining genes enriched in 2–5-bp indels. **f**, SNV and indel rates (total events per Mb per patient) in different functional regions of 18 protein-coding genes enriched in 2–5-bp indels (without *ALB*, which contributed 47% of indels). Red lines indicate background indel and SNV rates estimated from all protein-coding genes. Error bars as in **d**; raw counts provided in Supplementary Table 18. **c–f**, Mutations analysed in all unique cases ( $n = 2,583$ ).

that associates with *ERBB2*, and also affects migration and invasion in gastric cancer<sup>23</sup>. *TOB1* regulates other mRNAs through binding to their 3' UTR and promoting deadenylation<sup>24</sup>. Tumours with 3' UTR mutations in *TOB1* showed a trend towards decreased expression ( $P = 0.053$ , fold difference = 0.7). The mutations did not concentrate in known miRNA-binding sites; however, the region is extremely conserved and thus probably functional (Fig. 2a). *TOB1* and its neighbouring gene *WFIKN2* are focally amplified in breast cancer and pan-cancer, suggesting a complex role in cancer (Extended Data Fig. 4b). *NFKBIZ* is a transcription factor that is mutated in diffuse large B cell lymphoma and amplified in primary lymphomas<sup>25</sup>. Mutations in the 3' UTR accumulated in a hotspot proximal to the stop codon and upstream of conserved miRNA-binding sites (Extended Data Fig. 5b). The enrichment of indels next to the stop codon suggests that this hotspot is not due to AID off-target activity. Previous functional experiments have associated these mutations with increased *NFKBIZ* expression<sup>25</sup>, which we observed in our lymphoma cohort ( $P = 0.035$ , fold difference = 3.2; after correction for copy number,  $P = 0.03$ ) (Extended Data Fig. 5b).

Both the exon and promoter of the non-coding RNA *RMRP* were significantly mutated in multiple types of cancer (Fig. 1b, Extended Data Fig. 5c). Germline *RMRP* mutations cause cartilage–hair hypoplasia, and previous in vitro studies have shown that some somatic promoter mutations are functional<sup>16</sup>. The *RMRP* locus is also focally amplified in several types of tumour (Extended Data Fig. 4b). The enrichment of mutations in sites that can affect secondary structure suggests that these mutations are functional ( $P = 0.011$ , permutation test) (Extended

Data Fig. 5c), although caution is required because this locus also appears to be affected by mapping artefacts or increased mutation rates (Supplementary Note 4).

The miR-142 precursor miRNA was significant in Lymph–BNHL and the lymphatic and haematopoietic cohorts (Fig. 1b; Extended Data Fig. 5d). The locus is a known AID off-target region in lymphoma<sup>12,26</sup>, but 7 out of 8 mutations in the mature miRNA *mir-142-p3*—for which the largest functional effect is expected—were not assigned to AID, which suggests that these mutations are under selection<sup>12</sup>.

## Unbiased genome-wide driver screen

To test whether we missed drivers by focusing on functionally annotated regions, we applied an unbiased genome-wide survey to all non-overlapping 2-kb windows for excess point mutations. Twenty-two of the resulting 67 significant windows overlap with known protein-coding drivers, and 28 overlap highly transcribed regions with an excess of 2–5-bp indels (described in the 'Transcription-associated indel signature' section below) (Extended Data Fig. 5e, Supplementary Table 9, Supplementary Note 5). The remaining 17 windows have no obvious link to cancer, and several appear to be affected by mapping artefacts. A separate analysis of 4,351 ultra-conserved non-coding regions did not yield new candidate drivers (Extended Data Fig. 5e, Supplementary Note 5). Both screens suggest that the paucity of non-coding point-mutation drivers found in this study is not due to the annotation of functional elements.

## Increasing power for known cancer genes

Finally, we performed restricted hypothesis testing to boost the statistical power to detect *cis*-regulatory driver mutations near cancer genes from the CGC<sup>21</sup> (Supplementary Table 7). Restricted hypothesis testing of cancer gene promoters revealed a significant recurrence of *TP53* promoter mutations (11 patients in pan-cancer,  $Q = 0.044$ ), mostly comprising SNVs and deletions that affect the transcription start site or donor splice site of the first non-coding exon. In 10 out of 11 cases, the mutation occurred in combination with loss-of-heterozygosity, and all samples with expression data showed decreased mRNA levels (Fig. 2b). None of these patients contained additional coding mutations that could instead be responsible for the downregulation of *TP53*. To our knowledge, this is the first report of a relatively infrequent—but impactful—form of *TP53* inactivation by non-coding mutations.

Focal gains or losses in cancer are selected for modulating expression levels of their target genes. Restricting the hypothesis testing to the non-coding elements of such genes ( $n = 216,986$  cohort–element combinations, representing 5,201 unique elements) (Methods) yielded only one new hit, the 3' UTR of the oncogene *FOXAI* in prostate cancer (Supplementary Table 11).

## Transcription-associated indel signature

Several significant non-coding elements (the *ALB* 3' UTR, *NEAT1*, *MALAT1* and *MIR122*) were hit by many indels; all have previously been reported to be mutated in cancer<sup>10,15,27</sup> (Figs. 1b, 2c). To explore whether *ALB* 3' UTR events are under selection, we calculated indel rates across the functional regions of this gene. The indel rate is notably high throughout the UTRs, introns and exons, and even downstream of the polyadenylation site—a pattern inconsistent with selection (Fig. 2c, d). Similarly, *FOXAI* has high indel rates throughout its locus, whereas the indels in *NFKB1Z* and *TOB1* are in their 3' UTRs, suggesting that these are driver events (Fig. 2d). *ALB*, *NEAT1* and *MALAT1* mutations were not associated with changes in gene expression (Extended Data Fig. 4a) and were not associated with high cancer cell fractions or biallelic loss (Extended Data Fig. 6a, b). Likewise, indels in *MIR122* were downstream of the mature miRNA, and were not associated with altered expression of the targets of this miRNA (Supplementary Note 5).

If the indels in these genes were due to a mutational process rather than selection, they might exhibit distinct features. Indeed, indels in *NEAT1*, *MALAT1*, *MIR122* and *ALB* were strongly enriched in 2–5-bp-long events (Fisher's  $P < 6.8 \times 10^{-5}$ , for all) (Fig. 2e). A systematic search of coding and non-coding genes with significantly ( $Q < 0.1$ ) increased rates of 2–5-bp indels revealed that this mutational process affects at least 18 additional genes in different types of tumour, most of which are highly expressed and tissue-specific (as has previously been reported for some of these genes<sup>15</sup>) (Extended Data Fig. 6e, f). Although less enriched, SNVs also occur at high frequencies in these regions (Fig. 2f). Overall, our findings suggest that the indels in *MALAT1*, *NEAT1*, *ALB* and *MIR122* are not driver events and are the result of a transcription-associated mutational process. The previously reported oncogenic effect of altered *MALAT1* and *NEAT1* expression<sup>27–29</sup> may thus be unrelated to these mutations. Our findings also suggest that although *FOXAI* protein-coding indels are drivers, 3' UTR indels might be passengers<sup>30</sup>.

## Breakpoints at driver and fragile sites

Driver structural variants may act by disrupting one or both of their breakpoint loci (for example, deactivating a tumour suppressor), or by generating a novel juxtaposition between loci. We thus searched both for genomic regions with SRBs and for pairs of regions with SRJs (Extended Data Fig. 7).

For SRBs, we first defined a background model to predict breakpoint density, using eight explanatory variables (Methods, Supplementary

Table 13) and accounting for unexplained sources of variation<sup>31</sup> (Supplementary Note 6). We identified 53 disjoint regions with SRBs ( $Q < 0.1$ ) (Fig. 3a, Supplementary Table 14), which cleanly divided into two groups on the basis of the variability of the breakpoints at the other side of the rearrangements. Eight SRBs had partner breakpoints that were tightly clustered (had low rearrangement dispersion scores; Methods) and represented known oncogenic fusions. The remaining 45 SRBs had dispersed partner breakpoints (had high rearrangement dispersion scores), and were largely associated with previously identified somatic copy-number alterations (SCNAs) (Fig. 3b).

It has been difficult to distinguish recurrent driver SCNAs from passenger events at fragile sites<sup>32</sup>. At the resolution afforded by whole-genome sequencing, late replication timing predicted fragility-associated SRBs better than existing fragile site annotations (Supplementary Note 7), identifying 12 fragile-like SRBs (Fig. 3b). The remaining 33 SCNA-like SRBs comprised 14 amplifications, 8 deletions and 11 copy-neutral events (Supplementary Table 14).

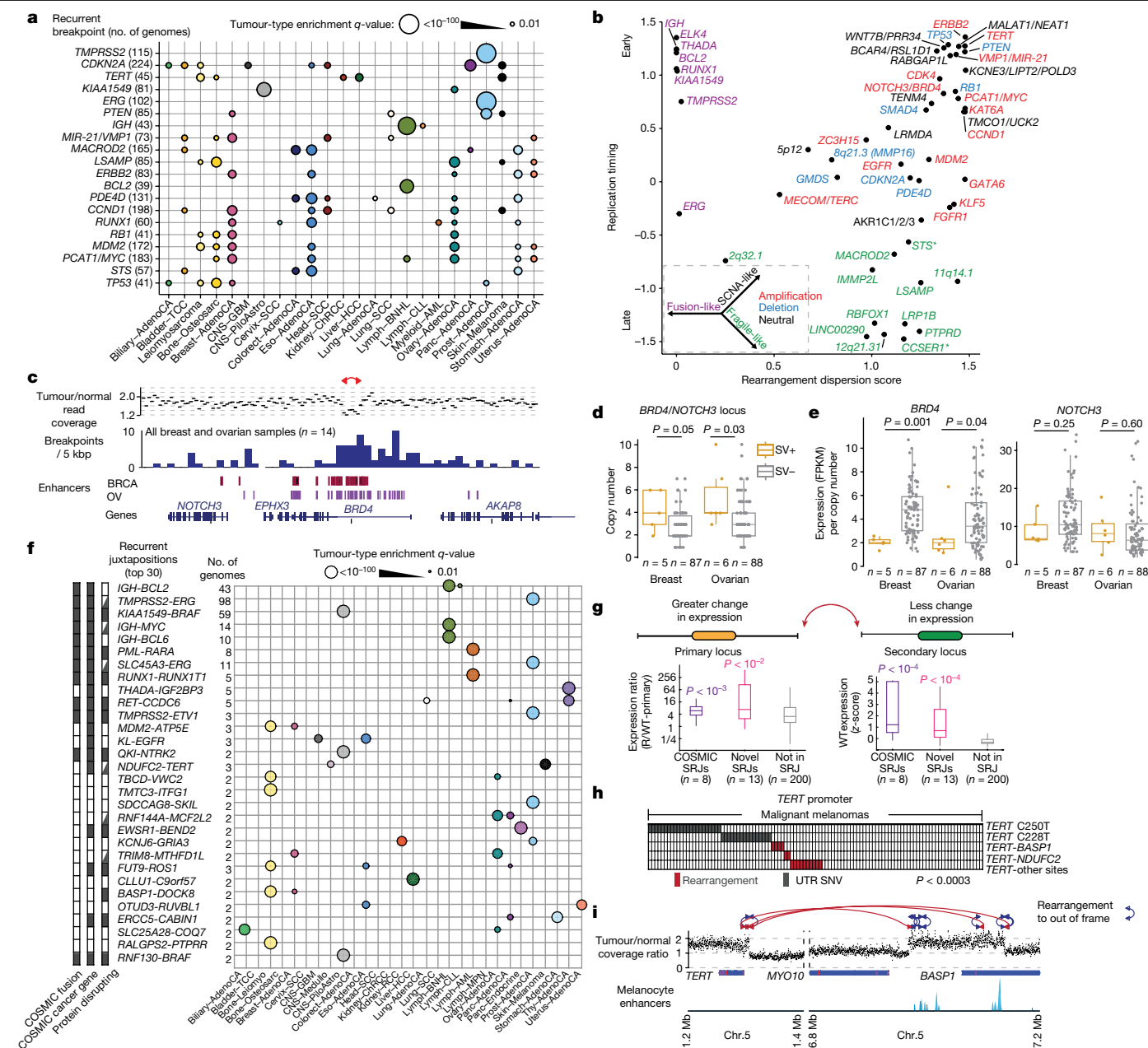
The different classes of SRB were associated with different effects on neighbouring genes. Five of the eight deletion-associated SRBs were associated with biallelic inactivation of nearby known tumour suppressors, compared to none of the 12 fragile-like SRBs ( $P = 0.039$ ) (Extended Data Fig. 8a). The fragile-like SRBs were furthest from tissue-matched enhancers and caused the weakest expression changes, consistent with them being passenger events<sup>32</sup>. By contrast, fusion-like SRBs were closer to tissue-matched enhancers than the other SRBs ( $P < 0.01$ ) (Extended Data Fig. 8b) and were associated with greater changes in expression than all other SRBs except amplifications ( $P < 0.05$  for all types) (Extended Data Fig. 8c, Methods). Our analyses indicate that SRB driver events can be classified using rearrangement dispersion scores, replication timing and gene expression. Notably, neither rearrangement dispersion scores nor association with replication time can be accurately determined from microarrays or whole-exome sequencing, which highlights the importance of whole-genome sequencing. Altogether, we identified SRBs at 34 sites of known oncogenic fusions and recurrent SCNAs, 5 additional sites that are probably due to DNA fragility and 14 novel driver candidates (Supplementary Note 8).

## Novel structural-variant driver candidates

Although most SCNA-like SRBs act by altering gene copy numbers, several appeared to target regulatory elements. We identified three that were significantly ( $Q < 0.05$ ) associated with expression changes of nearby genes after controlling for copy number (Methods), two of which we discuss here. The first comprised structural variants at 10p15, which were associated with a greater than twofold upregulation of *AKRIC1*, *AKRIC2* and *AKRIC3* in seven cases of lung squamous cell carcinoma and two cases of liver hepatocellular carcinoma (Extended Data Fig. 8d). AKRIC proteins are aldo-keto reductases involved in steroid homeostasis. Ectopic expression transforms cell lines, and germline mutations have previously been linked to an increased risk of developing lung cancer<sup>33,34</sup>. Three-quarters of the breakpoints are near (<10 kb) lineage-specific enhancers, potentially altering promoter–enhancer interactions (and hence gene expression). However, because the highest density of breakpoints lies between two long inverted repeats, the structural variants may have been induced by DNA secondary structure.

The second SRB contains recurrent microdeletions (<50 kb) involving the 5' end of *BRD4* in ovarian (eight cases,  $P < 10^{-7}$ ) and breast tumours (six cases,  $P < 0.04$ ) (Fig. 3c, Extended Data Fig. 8e). These deletions were highly enriched in cancers that amplified a segment that includes *BRD4* and *NOTCH3* ( $P < 0.004$ ) (Fig. 3d, Extended Data Fig. 8f) but were not a direct consequence of these amplifications (Supplementary Note 9). *BRD4* is a chromatin regulator and a therapeutic target in several types of cancer<sup>35,36</sup>, including ovarian and triple-negative breast cancer<sup>37,38</sup>. Given the increased copy number of the full *BRD4* gene, we would expect increased gene expression. However, the microdeletions





**Fig. 3 | Significantly recurrent breakpoints and juxtapositions.** **a**, Relative enrichment (Fisher's exact test) for events per tumour type for the 20 most significant SRBs (circle size). Loci are labelled by the likely driver gene from the CGC<sup>21</sup>. For gene symbols separated by a solidus, both or either of the genes are intended. **b**, Rearrangement dispersion score versus mean replication timing of the 53 SRBs. Colours indicate fusion (purple), fragile-like (green), deletion (blue), amplification (red) or copy-neutral (black) events. **c**, Tumour-to-normal read coverage ratio in an ovarian tumour with a *BRD4* microdeletion; red arrow indicates the rearrangement (top). Breakpoint density across PCAWG breast and ovarian cancers (middle). Enhancer locations from breast (BRCA) and ovarian (OV) tissue<sup>51</sup> (bottom). **d**, Somatic copy number at the *BRD4* and *NOTCH3* locus in breast and ovarian cancers with (SV+) and without (SV-) rearrangements. **e**, Gene expression per absolute copy number for *BRD4* and *NOTCH3*. **f**, The 30 most significant SRJs, with their relative enrichment (circle size) per tumour type, annotated with oncogenic fusions from the Catalogue of Somatic Mutations in Cancer (COSMIC) (left), CGC gene (centre) and protein disruption (right) (Methods). *ATP5E* is also known as *ATP5F1E*. **g**, Expression

are associated with a lower expression of *BRD4* in breast ( $P = 0.001$ ) and ovarian tumours ( $P = 0.04$ ), but not of the neighbouring gene *NOTCH3* (Fig. 3e). The focal deletions in *BRD4* overlap a prominent

correlates of rearrangements in SRJs from COSMIC (purple), other SRJs (pink) or not in any SRJ (grey). For each rearrangement (R), the primary locus (left) is defined as the breakpoint within 100 kb of the gene that is most overexpressed in rearranged samples; the secondary locus (right) is the other breakpoint. Expression at the primary locus in samples with the rearrangement relative to samples without the rearrangement is greater for SRJs than for other rearrangements (left). The tissue-specific expression at the secondary locus in wild-type (WT) samples, relative to samples of different tissue types, is greater for SRJs than other rearrangements (right).  $P$  values represent comparisons to 'not in SRJ'. **d**, **e**, **g**, Box plots show the interquartile range, median and 95% confidence interval; two-sided  $t$ -test. **h**, *TERT* promoter mutations and rearrangements across PCAWG melanomas. **i**, Rearrangements between *TERT* promoter and *BASP1* and *MYO10* locus result in focal amplification and relocation of distal enhancers to *TERT*. AML, acute myeloid leukaemia; Colorect, colorectal; Leiomyo, leiomyosarcoma; MPN, myeloproliferative neoplasm; Osteosarc, osteosarcoma; PiloAstro, pilocytic astrocytoma.

exon-1 H3K4me3 peak and intron-1 enhancer elements in HMEC (normal breast) and MCF-7 (breast tumour) cells (Extended Data Fig. 8e), which suggests that these deletions disrupt regulatory elements.

To our knowledge, this is the first evidence of a recurrent microdeletion limiting expression of an amplified gene.

## Recurrent fusions target gene regulation

Motivated by the detection of fusion-like SRBs, we specifically looked for genomic loci that were juxtaposed more often than expected by chance, after controlling for both the rate of breakpoints at each locus and the distance between them (Methods). We identified 90 such SRJs (Fig. 3f, Supplementary Table 15), including 13 known oncogenic fusions (including all 8 fusion-like SRBs) and 77 novel hits—18 of which linked to at least one known cancer gene (Supplementary Note 8). Previously reported oncogenic SRJs were observed more frequently (average 24 patients per fusion, range 2–98) than novel ones (most often 2 patients per fusion, range 2–4). As juxtapositions are unlikely to occur by chance, observing even two becomes highly significant. However, it is possible that some SRJs reflect inaccuracies in our background model rather than true drivers. We therefore further evaluated the SRJs on the basis of (i) a ‘robustness factor’ that indicates how much the background rate could increase before the SRJ would become insignificant, and (ii) the ratio between the observed and expected numbers of events under the current background model (‘effect size’) (Extended Data Fig. 9a). Twenty-six SRJs, including 11 of the 13 known drivers and 15 newly identified SRJs, are robust to tripling the expected background rate, and 22 others would remain significant with a doubled rate.

Most canonical driver rearrangements have previously been found in single tumour types, often associated with tissue-specific expression<sup>39,40</sup>. We found that 9 of our top 10 SRJs are tissue-specific, despite searching across 30 different types of tumour. Such tissue specificity is not observed for cancer genes affected by SCNAs, for which the top 10 are altered in 11.9 cancer types (on average), or by point mutations (for which the top 10 are altered in 6.7 cancer types, on average) (Supplementary Table 16).

The tissue specificity of SRJs suggests that they are strongly shaped by epigenetic state, either owing to mechanistic reasons (for example, tissue-specific three-dimensional proximity of the two DNA breakpoints) or to selection that connects tissue-specific regulatory elements with oncogenes<sup>13,41–43</sup>. The latter seems to be more likely because: (i) SRJs are associated with significant overexpression of only one of the rearrangement partners (the ‘primary locus’) relative to randomly selected rearrangements (primary locus,  $P < 10^{-4}$  (Fig. 3g left); secondary locus,  $P > 0.05$  (Extended Data Fig. 9b left)); (ii) the rearrangement partner, in the secondary locus, tends to be highly expressed in that tissue type relative to others (Fig. 3g right); and (iii) the distance to the nearest tissue-specific enhancer is smaller for SRJs than for rearrangements overall (Extended Data Fig. 9b). These observations suggest that SRJs act in general by bringing regulatory elements to an oncogene that is otherwise expressed at a low level.

In many cases, SRJs generate truncated or chimeric proteins, and breakpoints within introns or exons were indeed overrepresented (68% versus 56% expected,  $P < 10^{-7}$ ). However, only 11 of the 30 (37%) most significant SRJs generated novel proteins in all samples, and 6 others sometimes generated novel proteins; the rest were either non-disruptive or contained breakpoints within the first two introns of the disrupted gene, leaving most of the protein intact<sup>44</sup> (Fig. 3f). Moreover, SRJs that generate novel proteins exhibited expression changes similar to those that do not ( $P = 0.4$ ) (Extended Data Fig. 9c). We conclude that altering gene expression is a key function of both classes of SRJs, and that SRJs are akin to non-coding driver point mutations that act on regulatory elements.

We found several SRJs that involve amplified oncogenes, including *MDM2*, *EGFR* and *TERT* (Fig. 3f, h, i, Extended Data Fig. 9d–f, Supplementary Table 15). The *TERT* promoter region was juxtaposed in four melanomas ( $P < 10^{-7}$ ) to a region in the *BASPI* gene (both on chromosome 5), and to a region near *NDUFC2* (t(5,11)) in two melanomas and

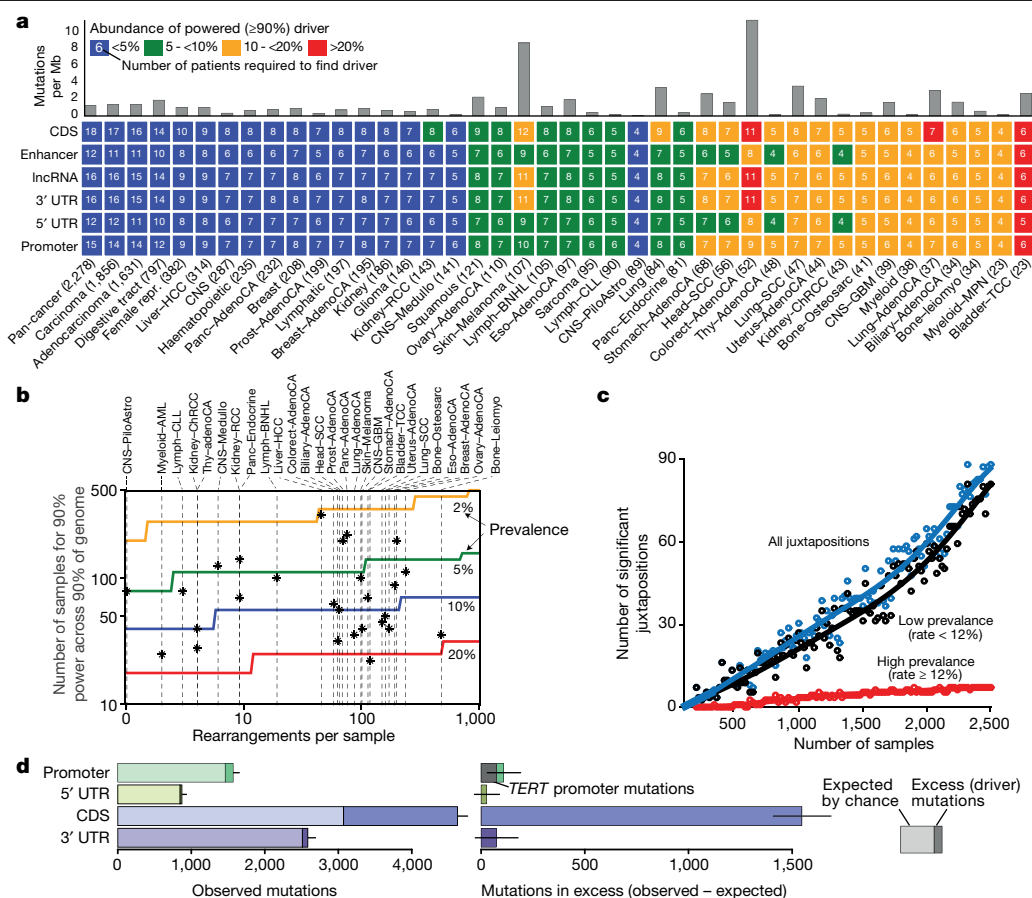
one medulloblastoma ( $P < 10^{-8}$ ). Both juxtaposed regions were marked with melanocyte enhancers, which suggests that they could drive *TERT* expression. Among melanomas, these rearrangements are mutually exclusive with the C228T and C250T mutations of the *TERT* promoter ( $P < 10^{-3}$ ) (Fig. 3h). Because the juxtapositions were always part of complex events that also amplified *TERT*, increased *TERT* expression may be due to amplification, the juxtapositions or both.

## Paucity of non-coding drivers in cancer

Our analyses of genomic hotspots, functional elements, genomic windows and SRJs all suggest that non-coding drivers are rare compared to protein-coding drivers. This might, in part, be due to a lack of discovery power<sup>3</sup>. We therefore evaluated the discovery power of mutational-burden tests for recurrent events across the different types of element in our tumour cohorts, focusing first on point mutations<sup>3,16</sup>. We found that the fraction of mutated patients required for a driver to reach 90% discovery power ranged from <1% in large cohorts with low background-mutation densities to 25% in small cohorts with high background-mutation densities (Fig. 4a). Different types of element were similarly powered, suggesting that the paucity of drivers in non-coding versus coding elements is not due to a lack of power. Similarly, our power to detect SRJs was higher in large cohorts with low rearrangement rates, and for long and interchromosomal rearrangements owing to their lower overall rates (Extended Data Fig. 10a): we were only powered to detect events that recur in 5–20% of samples in most types of cancer (Fig. 4b). Moreover, beginning with about 2,500 tumours, we expect to find a new SRJ with every 25 additional genomes (Fig. 4c).

Low sequencing coverage (for example, in GC-rich regions<sup>45</sup>) also limits driver discovery. To measure this effect in the PCAWG data, we quantified our ability to detect mutations (detection sensitivity)<sup>16</sup> in cancer gene promoters. Although the mean detection sensitivity in promoters is high (41.9% of genomic positions have mean detection sensitivity >80% across tumours), only 4.1% of the promoters had detection sensitivity >90% in >90% of bases. In particular, the two canonical *TERT* promoter hotspots had highly variable detection sensitivity among patients and cohorts, from only 3% of patients in the central-nervous-system pilocytic astrocytoma cohort to 100% in the thyroid adenocarcinoma cohort (Extended Data Fig. 10b). From these data, we inferred the expected number of *TERT* events in each tumour type (Extended Data Fig. 10c) and found that about 263 (95% confidence interval 232–295) *TERT* hotspot mutations were probably missed owing to a lack of detection sensitivity. Moreover, on average 9.9% (1.3–13.0% interquartile range) of the cancer gene promoter territory in the tumour of each patient was severely underpowered (an average detection sensitivity of <10%). Therefore, the lack of coverage in promoters may contribute to the paucity of non-coding drivers.

To determine whether the paucity of non-coding drivers discovered thus far could be due to the limited statistical power of current datasets, we estimated the overall excess of point mutations above background (that is, the expected number of driver events) in coding and *cis*-regulatory non-coding sequences in 603 cancer genes<sup>46</sup> (Methods, Supplementary Table 7, Supplementary Note 11). To minimize the effect of samples with low detection sensitivity, we included only 936 samples with >90% detection sensitivity at the two *TERT* promoter hotspots (Extended Data Fig. 10c, d, Supplementary Note 11). Overall, this approach predicted more than 1,475 driver mutations (95% confidence interval 1,410–1,687; 1,069 SNVs and 406 indels) in the protein-coding sequences of these cancer genes (Fig. 4d), compared to only 96 (95% confidence interval 30–190) estimated driver mutations in promoters (73 attributed to *TERT*), 22 (95% confidence interval 0–88) in 5'UTRs, and 68 (95% confidence interval 0–178) in 3'UTRs. Non-coding mutations in cancer-gene promoters were also not generally associated with loss-of-heterozygosity or altered expression, as one would expect if they were enriched with drivers (Supplementary Note 12).



**Fig. 4 | Power considerations and paucity of non-coding drivers.** **a**, Heat map shows the minimal frequency of a driver element with  $\geq 90\%$  discovery power. Power is dependent on the background mutation frequency (above the heat map), the element length (median length depicted in Extended Data Fig. 2c) and the number of patients with mutations (cell numbers). For example, the pan-cancer cohort is powered to discover a protein-coding driver gene (coding sequence (CDS)) present in  $<1\%$  (18 patients), whereas the Bladder-TCC cohort is only powered to discover drivers present in at least 27% (6 patients). **b**, Number of samples required to detect 90% of recurrent juxtapositions across 90% of pairs of loci, as a function of the median number of rearrangements per sample and the rate above background at which the fusion recurs (solid lines). The vertical dashed lines represent the median

rearrangement rates of each cancer type, and the stars on these lines indicate the numbers of whole genomes analysed for that cancer type. **c**, Number of SRJs detected after downsampling the data to various sample sizes, separately indicating rearrangements that recur at high ( $\geq 12\%$ ; red) and low ( $< 12\%$ ; black) rates above background; their sum (blue). **d**, Number of observed mutations (SNVs and indels) in *cis*-regulatory and coding regions of 603 protein-coding cancer genes with the expected numbers shown in lighter colours (left). Right, the number of excess mutations (that is, the estimated number of driver mutations) (right). The grey fraction of promoter mutations indicates *TERT* events. Error bars show 95% binomial confidence intervals. Only samples with high detection sensitivity were included ( $n = 936$ ).

These results collectively indicate that, independently of statistical power, non-coding *cis*-regulatory driver mutations in known cancer genes besides *TERT* are much less frequent than protein-coding drivers.

## Discussion

The accurate and reliable discovery of genomic drivers in tumours may have critical implications for patients with cancer. Our findings and the methods introduced here for the discovery of point-mutation and structural-variant drivers, method integration, vetting of candidates and identification of local hypermutation and fragile sites represent an important contribution to the collective effort towards charting all malignant changes that drive the cancer of each patient<sup>5</sup>.

Among the most interesting candidate non-coding driver elements we uncovered are the 5'-end mutations in *TP53*; 3' UTR mutations in *NFKB1* and *TOB1*; and rearrangements involving *AKR1C* genes and *BRD4*. By careful analysis of the whole-genome sequencing data, we found that several previously reported and frequently altered non-coding elements may not be genuine drivers, including (i) the non-coding RNAs, *NEAT1* and *MALAT1* (which contain a high density of indels,

seemingly owing to a transcription-associated mutational process) and (ii) recurrent structural variants in regions of late replication, indicating DNA fragility.

This study yielded unexpectedly few non-coding driver point mutations and structural variants. SRJs, which appear to act largely through the rearrangement of regulatory elements, are less frequent than SCNA-like SRBs, which directly amplify or delete coding sequences. The results from five analyses—hotspot recurrence, driver-element discovery, structural variants, discovery power and aggregated mutational excess—suggest that this paucity is not caused by a particular analysis strategy, but that regulatory elements truly contribute a much smaller number of recurrent cancer-driving events than protein-coding sequences. This paucity of non-coding drivers contrasts with the distribution of germline polymorphisms associated with heritability of complex traits, which are most frequently located outside of protein-coding genes<sup>47</sup>.

At least two factors contribute to the relative paucity of non-coding driver mutations in cancer: (i) the differential fitness effects of coding and non-coding mutations and (ii) the target size of functional elements. The paucity of promoter driver mutations in well-established

cancer genes suggests that point mutations markedly affect the function of non-coding regulatory elements only rarely. This highlights *TERT* as a notable exception, perhaps because even a modest increase in *TERT* expression may suffice to circumvent normal telomere shortening. For other cancer genes, directly mutating protein-coding sequences or altering expression levels by copy-number change may provide larger phenotypic effects. For example, complete loss-of-function by nonsense mutations or deletions may be easier to achieve than by disrupting or translocating regulatory regions.

Technical shortcomings (such as coverage ‘blind spots’ in GC-rich promoters and different filtering strategies) may cause genuine drivers to be missed<sup>48</sup>. Therefore, the discovery of non-coding drivers will benefit from technical improvements, including even sequence coverage, longer and accurate reads, and improved variant-calling methods. Moreover, better annotation of functional non-coding elements will increase both the power to discover infrequently mutated driver elements and their interpretability. As datasets grow, yet-unidentified mutational mechanisms targeting particular genomic regions will emerge and require improved background models, including additional covariates and more-sophisticated statistical models. The analysis of structural variants has greater challenges because (i) accurately modelling their background density is complicated by their lower frequency and larger fraction of drivers (Supplementary Note 6); (ii) their target genes may be far from the breakpoints, as in SCNAs; (iii) the space for modelling SRJs is much larger (the genome squared); and (iv) many structural variants are part of complex events that often involve multiple chromosomes<sup>31</sup>, so that the resultant topology cannot be deduced without technologies such as long- or linked-read sequencing<sup>49,50</sup>. For these reasons, experimental validation remains important for all—and especially for non-coding—candidate drivers.

Our work suggests that larger datasets and technological advances will continue to identify new non-coding drivers, albeit at considerably lower frequencies than protein-coding drivers. We anticipate that the approaches developed here will provide a solid foundation for the incipient era of driver discovery from ever-larger numbers of cancer whole genomes.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1965-x>.

- Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **174**, 1034–1035 (2018).
- Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
- Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
- Beroukhi, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
- Horn, S. et al. *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
- Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
- Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
- Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
- Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
- Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
- Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
- Northcott, P. A. et al. Enhancer hijacking activates GF11 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
- Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
- Imielinski, M., Guo, G. & Meyerson, M. Insertions and deletions target lineage-defining genes in human cancers. *Cell* **168**, 460–472.e14 (2017).
- Rheinbay, E. et al. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).
- Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **529**, 110–114 (2016).
- Perera, D. et al. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
- Sabarathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
- Mao, P. et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* **9**, 2626 (2018).
- Forbes, S. A. et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* **38**, D652–D657 (2010).
- Zhang, W. et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).
- Li, B.-S. et al. MicroRNA-25 promotes gastric cancer migration, invasion and proliferation by directly targeting transducer of ERBB2, 1 and correlates with poor survival. *Oncogene* **34**, 2556–2565 (2015).
- Hosoda, N. et al. Anti-proliferative protein Tob negatively regulates CPEB3 target by recruiting Caf1 deadenylase. *EMBO J.* **30**, 1311–1323 (2011).
- Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).
- Robbiani, D. F. et al. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol. Cell* **36**, 631–641 (2009).
- Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
- Ke, H. et al. NEAT1 is required for survival of breast cancer cells through FUS and miR-548. *Gene Regul. Syst. Bio.* **10** (Suppl 1), 11–17 (2016).
- Han, Y., Liu, Y., Nie, L., Gui, Y. & Cai, Z. Inducing cell proliferation inhibition, apoptosis, and motility reduction by silencing long noncoding ribonucleic acid metastasis-associated lung adenocarcinoma transcript 1 in urothelial carcinoma of the bladder. *Urology* **81**, 209.e1–209.e7 (2013).
- Annala, M. et al. Frequent mutation of the *FOXA1* untranslated region in prostate cancer. *Commun. Biol.* **1**, 122 (2018).
- Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* <https://doi.org/10.1038/s41586-019-1913-9> (2020).
- Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
- Chien, C.-W., Ho, I.-C. & Lee, T.-C. Induction of neoplastic transformation by ectopic expression of human aldo-keto reductase 1C isoforms in NIH3T3 cells. *Carcinogenesis* **30**, 1813–1820 (2009).
- Lan, Q. et al. Oxidative damage-related genes *AKR1C3* and *OGG1* modulate risks for lung cancer due to exposure to PAH-rich coal combustion emissions. *Carcinogenesis* **25**, 2177–2181 (2004).
- Filippakopoulos, P. et al. Selective inhibition of BET bromodomains. *Nature* **468**, 1067–1073 (2010).
- Dawson, M. A., Kouzarides, T. & Huntly, B. J. P. Targeting epigenetic readers in cancer. *N. Engl. J. Med.* **367**, 647–657 (2012).
- Shu, S. et al. Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer. *Nature* **529**, 413–417 (2016).
- Baratta, M. G. et al. An in-tumor genetic screen reveals that the BET bromodomain protein, BRD4, is a potential therapeutic target in ovarian carcinoma. *Proc. Natl Acad. Sci. USA* **112**, 232–237 (2015).
- Tomlins, S. A. et al. Recurrent fusion of *TPMRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
- May, W. A. et al. The Ewing's sarcoma *EWS/FLI-1* fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than *FLI-1*. *Mol. Cell Biol.* **13**, 7393–7398 (1993).
- Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates *IRS4* and *IGF2* in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
- Mani, R.-S. & Chinnaiyan, A. M. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat. Rev. Genet.* **11**, 819–829 (2010).
- Schneider, G., Schmidt-Suppran, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer* **17**, 239–253 (2017).
- St John, J., Powell, K., Conley-Lacombe, M. K. & Chinni, S. R. *TPMRSS2-ERG* fusion gene expression in prostate tumor cells and its clinical and biological significance in prostate cancer progression. *J. Cancer Sci. Ther.* **4**, 94–101 (2012).
- Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87–94 (2012).
- Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Shuai, S. et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* **574**, 712–716 (2019).
- Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).

50. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).
51. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

<sup>1</sup>The Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>2</sup>Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA, USA. <sup>3</sup>Harvard Medical School, Boston, MA, USA. <sup>4</sup>Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark. <sup>5</sup>Wellcome Trust Sanger Institute, Hinxton, UK. <sup>6</sup>Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA, USA. <sup>7</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>8</sup>Harvard University, Cambridge, MA, USA. <sup>9</sup>Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>10</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>11</sup>Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain. <sup>12</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>13</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>14</sup>Department for BioMedical Research, University of Bern, Bern, Switzerland. <sup>15</sup>Graduate School of Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland. <sup>16</sup>Department of Medical Oncology, Bern University Hospital, University of Bern, Bern, Switzerland. <sup>17</sup>Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>18</sup>Bioquant Center, Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg, Heidelberg, Germany. <sup>19</sup>Department of Cell Biology, Harvard Medical School, Boston, MA, USA. <sup>20</sup>cBio Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. <sup>21</sup>Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>22</sup>Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX, USA. <sup>23</sup>Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>24</sup>Department of Genitourinary Medical Oncology - Research, Division of Cancer Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>25</sup>Department of Urology, Icahn school of Medicine at Mount Sinai, New York, NY, USA. <sup>26</sup>Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. <sup>27</sup>Department of Radiation Oncology, Harvard Medical School, Massachusetts General Hospital, Boston, MA, USA. <sup>28</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. <sup>29</sup>Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. <sup>30</sup>Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University,

Uppsala, Sweden. <sup>31</sup>European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, UK. <sup>32</sup>Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark. <sup>33</sup>Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. <sup>34</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. <sup>35</sup>Division of Computational Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>36</sup>Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. <sup>37</sup>Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea. <sup>38</sup>Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland. <sup>39</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. <sup>40</sup>SBGD Inc, Cambridge, MA, USA. <sup>41</sup>Department of Haematology, University of Cambridge, Cambridge, UK. <sup>42</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. <sup>43</sup>Biotech Research & Innovation Centre (BRIC), The Finsen Laboratory, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. <sup>44</sup>Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. <sup>45</sup>Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>46</sup>Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>47</sup>The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. <sup>48</sup>Genetics and Genome Biology Program, SickKids Research Institute, Toronto, Ontario, Canada. <sup>49</sup>Department of Gene Technology, Tallinn University of Technology, Tallinn, Estonia. <sup>50</sup>Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. <sup>51</sup>New York Genome Center, New York, NY, USA. <sup>52</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>53</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>54</sup>Department of Biology and Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA, USA. <sup>55</sup>Department of Pathology and Laboratory Medicine, and Englander Institute for Precision Medicine, and Institute for Computational Biomedicine, and Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. <sup>56</sup>MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA. <sup>57</sup>Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan. <sup>58</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA. <sup>59</sup>Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. <sup>60</sup>Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. <sup>61</sup>Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. <sup>62</sup>Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, CA, USA. <sup>63</sup>Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. <sup>64</sup>Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan. <sup>65</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. <sup>66</sup>Department of Computer Science, Yale University, New Haven, CT, USA. <sup>67</sup>Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. <sup>68</sup>A list of members and their affiliations appears in the online version of the paper. <sup>69</sup>Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. <sup>70</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>71</sup>Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. <sup>72</sup>A list of members and their affiliations appears in the Supplementary Information. <sup>73</sup>These authors contributed equally: Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira. <sup>74</sup>These authors jointly supervised this work: Joachim Weischenfeldt, Rameen Beroukhi, Iñigo Martincorena, Jakob Skou Pedersen, Gad Getz. \*e-mail: joachim.weischenfeldt@bric.ku.dk; rameen@broadinstitute.org; im3@sanger.ac.uk; jakob.skou@clin.au.dk; gadgetz@broadinstitute.org



PCAWG Drivers and Functional Interpretation Working Group

Federico Abascal<sup>15</sup>, Samirkumar B. Amin<sup>21,22</sup>, Gary D. Bader<sup>75</sup>, Pratiti Bandopadhyay<sup>1,7</sup>, Jonathan Barenboim<sup>76</sup>, Rameen Beroukhim<sup>16,70</sup>, Johanna Bertl<sup>4</sup>, Keith A. Boroevich<sup>73</sup>, Søren Brunak<sup>77,78</sup>, Peter J. Campbell<sup>5,41</sup>, Joana Carlevaro-Fita<sup>14,15,16</sup>, Dimple Chakravarty<sup>24,25</sup>, Calvin Wing Yiu Chan<sup>17,26</sup>, Ken Chen<sup>79</sup>, Jung Kyoong Choi<sup>80</sup>, Jordi Deu-Pons<sup>81,82</sup>, Priyanka Dhingra<sup>28,29</sup>, Klev Diamanti<sup>30</sup>, Lars Feuerbach<sup>9</sup>, J. Lynn Fink<sup>83,84</sup>, Nuno A. Fonseca<sup>31</sup>, Joan Frigola<sup>81</sup>, Carlo Gambacorti-Passerini<sup>85</sup>, Dale W. Garsed<sup>86,87</sup>, Mark Gerstein<sup>38,42,66</sup>, Gad Getz<sup>1,2,3,71</sup>, Qianyun Guo<sup>32</sup>, Ivo G. Gut<sup>88,89</sup>, David Haan<sup>90</sup>, Mark P. Hamilton<sup>33</sup>, Nicholas J. Haradhvala<sup>1,2</sup>, Arif O. Harmanci<sup>91,92</sup>, Mohamed Helmy<sup>93</sup>, Carl Herrmann<sup>17,18</sup>, Julian M. Hess<sup>1</sup>, Asger Hobolth<sup>32</sup>, Ermin Hodzic<sup>94</sup>, Chen Hong<sup>9,26</sup>, Henrik Hornshøj<sup>4</sup>, Keren Isaev<sup>12,34</sup>, Jose M. G. Izarzugaza<sup>77</sup>, Rory Johnson<sup>14,16</sup>, Todd A. Johnson<sup>37</sup>, Malene Juul<sup>4</sup>, Randi Istrup Juul<sup>4</sup>, Andre Kahles<sup>35</sup>, Abdullah Kahraman<sup>36</sup>, Manolis Kellis<sup>1,56</sup>, Ekta Khurana<sup>28,29,60,61</sup>, Jaegil Kim<sup>1</sup>, Jong K. Kim<sup>95</sup>, Youngwook Kim<sup>37</sup>, Jan Komorowski<sup>30,38</sup>, Jan O. Korbe<sup>196,97</sup>, Sushant Kumar<sup>39</sup>, Andrés Lanzós<sup>14,15,16</sup>, Erik Larsson<sup>98</sup>, Michael S. Lawrence<sup>12</sup>, Donghoon Lee<sup>39</sup>, Kjong-Van Lehmann<sup>35</sup>, Shantao Li<sup>91</sup>, Xiaotong Li<sup>91</sup>, Zhao Lin<sup>1,8</sup>, Eric Minwei Liu<sup>28,29</sup>, Lucas Lochovsky<sup>42</sup>, Shaoke Lou<sup>91,99</sup>, Tobias Madsen<sup>4</sup>, Kathleen Marchal<sup>100,101</sup>, Iñigo Martincorena<sup>5</sup>, Alexander Martinez-Fundichely<sup>102,103,104</sup>, Yosef E. Maruvka<sup>1,2</sup>, Patrick D. McGillivray<sup>99</sup>, William Meyerson<sup>91,105</sup>, Ferran Muñoz<sup>82,106</sup>, Loris Mularoni<sup>100,11</sup>, Hidewaki Nakagawa<sup>57</sup>, Morten Muhligh Nielsen<sup>4</sup>, Marta Paczkowska<sup>76</sup>, Keunchil Park<sup>37</sup>, Kiejung Park<sup>107</sup>, Jakob Skou Pedersen<sup>4,32</sup>, Tirso Pons<sup>108</sup>, Sergio Pulido-Tamayo<sup>100,101</sup>, Benjamin J. Raphael<sup>88</sup>, Jüri Reimand<sup>12,34</sup>, Iker Reyes-Salazar<sup>106</sup>, Matthew A. Reyna<sup>109</sup>, Esther Rheinbay<sup>1,2,3</sup>, Mark A. Rubin<sup>59,60,61</sup>, Carlota Rubio-Perez<sup>82,106,110</sup>, S. Cenk Sahinalp<sup>94,111,112</sup>, Gordon Saksena<sup>1</sup>, Leonidas Salichos<sup>91,99</sup>, Chris Sander<sup>19,20</sup>, Steven E. Schumacher<sup>1,7</sup>, Mark Shackleton<sup>97,113</sup>, Ofer Shapira<sup>1,7</sup>, Ciyue Shen<sup>19,20</sup>, Raunak Shrestha<sup>111</sup>, Shimin Shuai<sup>12,13</sup>, Nikos Sidiropoulos<sup>43</sup>, Lina Sieverling<sup>9,26</sup>, Nasa Sinnott-Armstrong<sup>44</sup>, Lincoln D. Stein<sup>12,13</sup>, Joshua M. Stuart<sup>62</sup>, David Tamborero<sup>10,11</sup>, Grace Tiao<sup>1</sup>, Tatsuhiro Tsunoda<sup>23,63,64</sup>, Husen M. Umer<sup>30</sup>, Liis Uusküla-Reimand<sup>48,49</sup>, Alfonso Valencia<sup>83,114</sup>, Miguel Vazquez<sup>83,115</sup>, Lieven P. C. Verbeke<sup>101,116</sup>, Claes Wadelius<sup>50</sup>, Lina Wadi<sup>12</sup>, Jiayin Wang<sup>117,118,119</sup>, Joachim Warrell<sup>91,99</sup>, Sebastian M. Waszak<sup>97</sup>, Joachim Weischenfeldt<sup>43,69</sup>, David A. Wheeler<sup>65</sup>, Guanming Wu<sup>120</sup>, Jun Yu<sup>121</sup>, Jing Zhang<sup>39</sup>, Xuanping Zhang<sup>118,122</sup>, Yan Zhang<sup>91,123,124</sup>, Zhongming Zhao<sup>125</sup>, Lihua Zou<sup>126</sup> & Christian von Mering<sup>36</sup>

PCAWG Structural Variation Working Group

Kadir C. Akdemir<sup>79</sup>, Eva G. Alvarez<sup>127,128,129</sup>, Adrian Baez-Ortega<sup>130</sup>, Rameen Beroukhim<sup>16,70</sup>, Paul C. Boutros<sup>76,131,132,133</sup>, David D. L. Bowtell<sup>136,134</sup>, Benedikt Brors<sup>135,136,137</sup>, Kathleen H. Burns<sup>138</sup>, John Busanovich<sup>1,7</sup>, Peter J. Campbell<sup>5,41</sup>, Kin Chan<sup>39</sup>, Ken Chen<sup>79</sup>, Isidro Cortés-Ciriano<sup>140,141,142</sup>, Ana Dueso-Barroso<sup>83</sup>, Andrew J. Dunford<sup>143</sup>, Paul A. Edwards<sup>144,145</sup>, Xavier Estivill<sup>146,147</sup>, Dariush Etemadmoghadam<sup>86,87</sup>, Lars Feuerbach<sup>9</sup>, J. Lynn Fink<sup>83,84</sup>, Milana Frenkel-Morgenstern<sup>148</sup>, Dale W. Garsed<sup>86,87</sup>, Mark Gerstein<sup>38,42,66</sup>, Dmitry A. Gordenin<sup>149</sup>, David Haan<sup>90</sup>, James E. Haber<sup>54</sup>, Julian M. Hess<sup>1</sup>, Barbara Hutter<sup>135,150,151</sup>, Marcijn Imielinski<sup>51,55</sup>, David T. W. Jones<sup>152,153</sup>, Young Seok Ju<sup>90,154</sup>, Marat D. Kazanov<sup>155,156,157</sup>, Leszek J. Klimczak<sup>158</sup>, Youngil Koh<sup>158,160</sup>, Jan O. Korbe<sup>196,97</sup>, Kiran Kumar<sup>1,7</sup>, Eunjung Alice Lee<sup>161</sup>, Jake June-Koo Lee<sup>141,162</sup>, Yilong Li<sup>40,41</sup>, Andy G. Lynch<sup>144,145,163</sup>, Geoff Macintyre<sup>144</sup>, Florian Markowitz<sup>144,145</sup>, Iñigo Martincorena<sup>5</sup>, Alexander Martinez-Fundichely<sup>102,103,104</sup>, Matthew Meyerson<sup>143,164,165,166,167</sup>, Satoru Miyano<sup>168</sup>, Hidewaki Nakagawa<sup>57</sup>, Fabio C. P. Navarro<sup>99</sup>, Stephan Ossowski<sup>189,146,169</sup>, Peter J. Park<sup>141,162</sup>, John V. Pearson<sup>170,171</sup>, Montserrat Puiggròs<sup>83</sup>, Karsten Rippe<sup>172</sup>, Nicola D. Roberts<sup>41</sup>, Steven A. Roberts<sup>173</sup>, Bernardo Rodriguez-Martin<sup>127,128,129</sup>, Steven E. Schumacher<sup>1,7</sup>, Ralph Scully<sup>174</sup>, Mark Shackleton<sup>97,113</sup>, Nikos Sidiropoulos<sup>43</sup>, Lina Sieverling<sup>9,26</sup>, Chip Stewart<sup>1</sup>, David Torrents<sup>83,114</sup>, Jose M. C. Tubio<sup>45,46,47</sup>, Izar Villasante<sup>83</sup>, Nicola Waddell<sup>170,171</sup>, Jeremiah A. Wala<sup>1,6</sup>, Joachim Weischenfeldt<sup>43,69</sup>, Lixing Yang<sup>175</sup>, Xiaotong Yao<sup>51</sup>, Sung-Soo Yoon<sup>160</sup>, Jorge Zamora<sup>127,128,129,154</sup> & Cheng-Zhong Zhang<sup>52,53</sup>

<sup>75</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.  
<sup>76</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada.  
<sup>77</sup>Technical University of Denmark, Lyngby, Denmark.  
<sup>78</sup>University of Copenhagen, Copenhagen, Denmark.  
<sup>79</sup>University of Texas MD Anderson Cancer Center, Houston, TX, USA.  
<sup>80</sup>Korea Advanced Institute of Science and Technology, Daejeon, South Korea.  
<sup>81</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain.  
<sup>82</sup>Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain.  
<sup>83</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain.  
<sup>84</sup>Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, Queensland, Australia.  
<sup>85</sup>University of Milano Bicocca, Monza, Italy.  
<sup>86</sup>Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia.  
<sup>87</sup>Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia.  
<sup>88</sup>CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.  
<sup>89</sup>Universitat Pompeu Fabra, Barcelona, Spain.  
<sup>90</sup>Biomolecular Engineering Department, University of California, Santa Cruz, Santa Cruz, CA, USA.  
<sup>91</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA.  
<sup>92</sup>Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA.  
<sup>93</sup>The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada.  
<sup>94</sup>Simon Fraser University, Burnaby, British Columbia, Canada.  
<sup>95</sup>Research Core Center, National Cancer Centre Korea, Goyang-si, South Korea.  
<sup>96</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK.  
<sup>97</sup>Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.  
<sup>98</sup>Computational Biology Center, Memorial Sloan Kettering Cancer

Center, New York, NY, USA.  
<sup>99</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA.  
<sup>100</sup>Department of Information Technology, Ghent University, Ghent, Belgium.  
<sup>101</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium.  
<sup>102</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA.  
<sup>103</sup>Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA.  
<sup>104</sup>Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA.  
<sup>105</sup>Yale School of Medicine, Yale University, New Haven, CT, USA.  
<sup>106</sup>Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain.  
<sup>107</sup>Cheonan Industry-Academic Collaboration Foundation, Sangmyung University, Cheonan, South Korea.  
<sup>108</sup>Spanish National Cancer Research Centre, Madrid, Spain.  
<sup>109</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA.  
<sup>110</sup>Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.  
<sup>111</sup>Vancouver Prostate Centre, Vancouver, British Columbia, Canada.  
<sup>112</sup>Indiana University, Bloomington, IN, USA.  
<sup>113</sup>Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia.  
<sup>114</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.  
<sup>115</sup>Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway.  
<sup>116</sup>Department of Information Technology, Interuniversitair Micro-Electronica Centrum (IMEC), Ghent University, Ghent, Belgium.  
<sup>117</sup>The McDonnell Genome Institute, Washington University, St Louis, MO, USA.  
<sup>118</sup>School of Computer Science and Technology, Xian Jiaotong University, Xian, China.  
<sup>119</sup>School of Electronic and Information Engineering, Xian Jiaotong University, Xian, China.  
<sup>120</sup>Oregon Health & Sciences University, Portland, OR, USA.  
<sup>121</sup>Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, Hong Kong, China.  
<sup>122</sup>The University of Texas Health Science Center at Houston, Houston, TX, USA.  
<sup>123</sup>Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA.  
<sup>124</sup>The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH, USA.  
<sup>125</sup>School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.  
<sup>126</sup>Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA.  
<sup>127</sup>Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.  
<sup>128</sup>Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain.  
<sup>129</sup>The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain.  
<sup>130</sup>Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK.  
<sup>131</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.  
<sup>132</sup>Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada.  
<sup>133</sup>University of California Los Angeles, Los Angeles, CA, USA.  
<sup>134</sup>Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia.  
<sup>135</sup>National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany.  
<sup>136</sup>Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany.  
<sup>137</sup>German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.  
<sup>138</sup>Johns Hopkins School of Medicine, Baltimore, MD, USA.  
<sup>139</sup>Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada.  
<sup>140</sup>Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK.  
<sup>141</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.  
<sup>142</sup>Ludwig Center, Harvard Medical School, Boston, MA, USA.  
<sup>143</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA.  
<sup>144</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK.  
<sup>145</sup>University of Cambridge, Cambridge, UK.  
<sup>146</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.  
<sup>147</sup>Quantitative Genomics Laboratories (qGenomics), Barcelona, Spain.  
<sup>148</sup>The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel.  
<sup>149</sup>Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA.  
<sup>150</sup>German Cancer Consortium (DKTK), Heidelberg, Germany.  
<sup>151</sup>Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg, Germany.  
<sup>152</sup>Hopp Childrens Cancer Center (KiTZ), Heidelberg, Germany.  
<sup>153</sup>Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany.  
<sup>154</sup>Wellcome Sanger Institute, Hinxton, UK.  
<sup>155</sup>Skolkovo Institute of Science and Technology, Moscow, Russia.  
<sup>156</sup>A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia.  
<sup>157</sup>Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia.  
<sup>158</sup>Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA.  
<sup>159</sup>Center For Medical Innovation, Seoul National University Hospital, Seoul, South Korea.  
<sup>160</sup>Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea.  
<sup>161</sup>Division of Genetics and Genomics, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA.  
<sup>162</sup>Ludwig Center at Harvard, Boston, MA, USA.  
<sup>163</sup>School of Medicine and School of Mathematics and Statistics, University of St Andrews, St Andrews, UK.  
<sup>164</sup>Harvard Medical School, Boston, MA, USA.  
<sup>165</sup>Dana-Farber Cancer Institute, Boston, MA, USA.  
<sup>166</sup>Department of Medical Oncology, University Hospital, University of Bern, Bern, Switzerland.  
<sup>167</sup>Department of Pathology, The University of Melbourne, Melbourne, Victoria, Australia.  
<sup>168</sup>The Institute of Medical Science, The University of Tokyo, Tokyo, Japan.  
<sup>169</sup>Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany.  
<sup>170</sup>Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia.  
<sup>171</sup>Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, Queensland, Australia.  
<sup>172</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany.  
<sup>173</sup>Center for Reproductive Biology, School of Molecular Biosciences, Washington State University, Pullman, WA, USA.  
<sup>174</sup>Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, USA.  
<sup>175</sup>Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL, USA.

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Detailed methods are provided as Supplementary Methods.

### Dataset generation

Out of 2,955 samples, we selected 2,583 unique donor samples for SNV and indel driver-discovery analysis on the basis of SNV quality control (Supplementary Methods). We found that 110 additional myeloid–AML samples had robust structural variant calls despite SNV artefacts; we included these in structural variant analyses, for a total of 2,693 samples. For tumour-type cohort analyses, we used only cohorts with at least 20 patients. Tumour meta-cohorts were defined by cell type of origin or by organ system (for example, lung for lung adenocarcinoma and lung squamous cell carcinoma). A pan-cancer meta-cohort was created by combining all tumour cohorts except for Skin–Melanoma and lymphoid tumours (Supplementary Methods).

### Hotspot SNV analysis

We selected the 50 most-frequent SNV hotspots. These were analysed to identify known driver events; mutational signature biases related to sequence palindromes, immunoglobulin loci and so on; and potential artefacts, including regional mapping problems (Supplementary Methods).

### Mutational signatures

We performed de novo global-signature discovery and signature attributions with SignatureAnalyzer's Bayesian non-negative matrix factorization method<sup>52</sup>, based on 1,697 channels—including 1,536 pentanucleotide sequence contexts for single-base substitutions, 83 indel features, and 78 doublet-nucleotide substitution classes (Supplementary Methods).

### Definition of genomic elements

GENCODE v.19 (ref.<sup>53</sup>) and other genomic resources were used to define functional genomic elements, including protein-coding genes (CDS, splice sites, 5' UTR, 3' UTR and promoters), long non-coding RNAs (gene body, splice site and promoters), short RNAs, miRNAs and enhancers (Supplementary Methods).

### Candidate-driver-mutation identification methods and combination of results

We obtained results (*P* values) from 13 methods of driver discovery, including ActiveDriverWGS<sup>54</sup>, CompositeDriver, DriverPower<sup>55</sup>, dndscv<sup>46</sup>, ExInAtor<sup>56</sup>, LARVA<sup>57</sup>, MutSig tools<sup>3</sup>, NBR<sup>10</sup>, ncdDetect<sup>58</sup>, ncDriver<sup>59</sup>, OncodriveFML<sup>60</sup> and regDriver<sup>61</sup>. We integrated the results of all these methods using a custom framework based on a previously published method<sup>62</sup> for combining *P* values. Results from individual methods that showed large deviations from the expected uniform null distribution of *P* values were excluded. This approach was evaluated on real and simulated data. We controlled the FDR within each of the sets of tested genomic elements by concatenating all combined Brown's *P* values from across all tumour-type cohorts and applying the Benjamini–Hochberg procedure<sup>63</sup>. Cohort–element combinations with *Q* values < 0.1 were designated as significant hits, and combinations with  $0.1 \leq Q < 0.25$  as 'near significance'. Extensive details are provided in the Supplementary Methods. In addition, we tested for element-independent recurrence with the NBR method on 2-kb bins spanning the entire genome, and non-coding ultraconserved regions<sup>64</sup>.

### Post-filtering of driver mutation candidates

We applied stringent filters to discern positive selection from technical artefacts and mutational processes. We required at least three

mutations to be present in candidate elements, in at least three patients of the tested cohort; more than 50% of mutations in mappable regions; less than 50% of mutations in palindromic DNA; and less than 50% of mutations attributed to APOBEC activity. For lymphoid tumours and skin melanoma, we required that <35% and <50% of mutations were attributed to the AID and UV-light mutational signatures, respectively. The FDR was recalculated after post-filtering.

### Candidate driver structural-variant analyses

We applied separate analyses to detect recurrent structural variant breakpoints and recurrent juxtapositions. For each analysis, we first binned breakpoints, accepting only one breakpoint per sample per bin. We then determined which bins had more breakpoints than expected by chance (the SRB analysis), and which pairs of bins (or 'tiles') were joined by more rearrangements than expected by chance (the SRJ analysis).

### Candidate driver breakpoints

We calculated the background rate of breakpoints per bin based on a Gamma–Poisson model<sup>15</sup> that took into account genomic covariates, breakpoint counts normalized by the number of bases within each bin that had sufficient mappability to be eligible for breakpoint detection and accounted for an observed overdispersion of breakpoint counts that probably reflects unaccounted-for covariates (Supplementary Methods). We used the Gamma–Poisson model to calculate the *P* value for each bin (that is, the probability that each bin would exhibit the observed number of breakpoints (or greater) by chance alone), applying the Benjamini–Hochberg procedure<sup>63</sup> to correct for multiple hypotheses.

### Post-filtering of driver breakpoint candidates

We scored each recurrent breakpoint locus on the basis of the average replication timing of its breakpoints, and filtered those loci with scores >0.5 as probable fragile sites<sup>65</sup>.

### Candidate driver juxtapositions

We developed a background model to indicate the probability that two loci would be joined, taking into account the observed rate at which each locus underwent DNA breaks (from the breakpoint analysis), the distance between them and the propensity for these rearrangements to reflect a break followed by invasion versus two breaks that were then joined. We determined the probability that each tile would contain the observed number of rearrangements using a binomial test, followed by controlling for multiple hypothesis testing using the Benjamini–Hochberg procedure<sup>63</sup>.

### Gene-expression analyses

Gene-expression data were provided by the PCAWG Transcriptome Core Group<sup>66</sup>, and also generated using the same approach for an extended set of non-coding transcripts (Supplementary Methods).

### Additional evidence for selection

In addition to associations between mutations or structural variants and expression, we looked for signals of copy-number-alteration recurrence using the GISTIC2 algorithm<sup>67</sup>. We also tested whether driver candidates showed significantly higher frequency of loss-of-heterozygosity in mutated samples using Fisher's exact test. We calculated cancer allelic fractions using ploidy and tumour purity predictions from a previous publication<sup>68</sup>.

### Mutational process and indel enrichment

For every gene, we calculated the proportion of indels of length 2–5 bp out of the total number of indels. This proportion was compared to the genome background proportion using a binomial test. We also compared the indel rate per gene (not distinguishing by length) to the background. Both sets of *P* values were corrected with the FDR method.

## Power calculations

We estimated our power to discover driver elements mutated at a particular frequency in the population as previously described<sup>3,16</sup>, but solving for the lowest frequency for a driver element in the patient population that is powered ( $\geq 90\%$ ) for discovery. The calculation of this lowest frequency takes into account (i) the average background mutation frequencies for each cohort–element combination; (ii) the median length and average detection sensitivity for each element type and patient cohort size; and (iii) a global desired false-positive rate of 10%. The effect of element length is discussed in Supplementary Note 10, and details are provided in Supplementary Methods. Power calculations for detection of recurrent juxtapositions was performed similarly, except over a two-dimensional genomic fusion map divided into  $100 \times 100$ -kb tiles (Supplementary Methods). We performed this analysis first as a function of the distance between breakpoints (Extended Data Fig. 10a) and second as a function of the median number of rearrangements per sample, spanning values represented by histologies with more than 15 samples (Fig. 4b).

## Estimation of the number of mutations in non-coding regions of known cancer genes

NBR was used to estimate the background mutation rate expected across cancer genes, using a conservative list of 19,082 putative passenger genes as background and including as covariates the local mutation rate, gene expression and averaged copy-number states. The resulting model predicted the number of passenger SNVs and indels expected by chance. By aggregating the expected numbers over 603 known cancer genes from the CGC<sup>69</sup> (CGC v.80) (Supplementary Table 7), we compared the observed and expected numbers of mutations. For this analysis, we excluded samples with problems of low detection sensitivity (Supplementary Methods).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Data associated with this Article are available at <https://dcc.icgc.org/releases/PCAWG/drivers>. SRBs and SRJs are available at [www.svscape.org](http://www.svscape.org). A list of data files used for analyses in this paper is provided in Supplementary Table 20. Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC and TCGA PCAWG Consortium are described in an accompanying Article<sup>5</sup>, and are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access information that could potentially identify participants, such as germline alleles and the underlying sequencing data, researchers will need to apply to the TCGA data access committee via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC data access compliance office (<http://icgc.org/daco>) for the ICGC portion of the dataset. In addition, to access somatic single-nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

## Code availability

The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public

at <https://dockstore.org/search?search=pcawg> under the GNU General Public License v.3.0, which allows for reuse and distribution. Code for *P* value combination from multiple driver methods is available from [https://github.com/broadinstitute/getzlab-PCAWG-pvalue\\_combination/](https://github.com/broadinstitute/getzlab-PCAWG-pvalue_combination/). Power calculation methods are available from [https://github.com/broadinstitute/getzlab-PCAWG-power\\_calculations](https://github.com/broadinstitute/getzlab-PCAWG-power_calculations). Structural variant methods are located at <https://github.com/mskilab/fishHook>, <https://github.com/walaj/ginseng> and <https://github.com/walaj/SVsig>. Links to individual driver discovery methods are provided in the corresponding section of the Supplementary Methods.

52. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
53. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE pProject. *Genome Res.* **22**, 1760–1774 (2012).
54. Wadi, L., Uuskula-Reimand, L., Isaev, K. & Shuai, S. Candidate cancer driver mutations in super-enhancers and long-range chromatin interaction networks. Preprint at <https://www.biorxiv.org/content/10.1101/236802v1> (2017).
55. Shuai, S., Gallinger, S. & Stein, L. DriverPower: combined burden and functional impact tests for cancer driver discovery. Preprint at <https://www.biorxiv.org/content/10.1101/215244v1> (2017).
56. Lanzos, A. et al. Discovery of cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features. *Sci. Rep.* **7**, 41544 (2017).
57. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res.* **43**, 8123–8134 (2015).
58. Juul, M. et al. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *eLife* **6**, e21778 (2017).
59. Hornshøj, H. et al. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. *NPI Genom. Med.* **3**, 1 (2018).
60. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
61. Umer, H. M. et al. A significant regulatory mutation burden at a high-affinity position of the CTCF motif in gastrointestinal cancers. *Hum. Mutat.* **37**, 904–913 (2016).
62. Brown, M. B. 400: a method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987 (1975).
63. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
64. Dimitrieva, S. & Bucher, P. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* **41**, D101–D109 (2013).
65. Mrasek, K. et al. Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. *Int. J. Oncol.* **36**, 929–940 (2010).
66. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations in cancer. *Nature* <https://doi.org/10.1038/s41586-020-1970-0> (2020).
67. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
68. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* <https://doi.org/10.1038/s41586-019-1907-0> (2020).
69. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

**Acknowledgements** We thank the ICGC and TCGA PCAWG Network and the PCAWG steering committee for enabling this work, and for guidance throughout the study. We thank K. Kübler for assistance with meta-cohort generation and R. Heller for discussion on FDR. We are grateful to the PCAWG steering committee, M. Meyerson and E. S. Lander for helpful feedback, and M. Miller for editing this manuscript. Work in the Getz laboratory was partially funded by the GDA grants (NIH U24CA143845 and NIH U24CA210999), G.G.'s funds at the Broad Institute and MGH. G.G. is also partially supported by the Paul C. Zamecnik Chair in Oncology in MGH. J.S.P. was partially funded by Independent Research Fund Denmark (12-126439 and 7016-00379) and The Danish Cancer Society (R124-A7869). R.B. received funds from the National Institutes of Health (U54CA143798, R01CA188228, R35GM127029, and R01CA215489), the DFCI-Novartis Drug Discovery Program, the Pediatric Low Grade Astrocytoma Foundation, the Cure Starts Now Foundation and The Fund for Innovation in Cancer Informatics. J.W. was partly funded by Independent Research Fund Denmark (4183-00233B and 8020-00282B) and Danish Cancer Society (R147-Rp12977). N.L.-B. acknowledges funding from the European Research Council consolidator grant 682398) and Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, MINECO/FEDER, UE). We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

**Author contributions** This work was carried out by the PCAWG Drivers and Functional Interpretation Group based on data from the ICGC and TCGA PCAWG Network. All authors have had access to read and comment on the manuscript. E.R., M.M.N., F.A., J.A.W. and O.S. contributed equally. G.T., H.H., J.M.H. and R.I.J. contributed equally. J. Weischenfeldt, R.B., I.M., J.S.P. and G.G. jointly supervised this work. The full list of PCAWG Consortium members and their affiliations appears in the Supplementary Information. For the discovery of

point-mutation drivers, the following authors contributed: A.L., C. Hermann, C.W., D.A.W., E.K., E.M.L., E.R., G.G., G.T., H.M.U., I.M., J. Kim, J.R., J.S.P., K.A.B., K.D., K.I., L.M., L.U.-R., M.M.N., M.P.H., N.S.-A., N.J.H., P.D., P.J.C., R.J., S.B.A., T.A.J. and T.T. contributed and curated genomic annotations; C. Hermann., C.W.Y.C., I.M., M.K., S.B.A. and Y.E.M. contributed randomized mutational datasets for driver discovery; A.L., A.G.-P., A.H., D.L., D. Tamborero, E.K., E.M.L., E.R., H.H., I.M., J. Bertl, J.C.-F., J.M.H., J. Komorowski, J.R., J. Zhang, K.D., K.I., L.L., L.M., L.D.S., L.U.-R., L.W., M.B.G., M.J., N.L.-B., O.P., P.D., Q.G., R. Sabarinathan, S.K., S.S. and T.M. contributed driver methods and results. E.R., G.G., Z.L. and G.T. implemented results integration; A. Kahraman., C.v.M., G.T., H.H., I.M., J.R., L.F. and M.M.N. contributed driver results integration; and C. Hermann, C.W.Y.C., E.K., E.R., G.G., J. Kim, J.M.H., J.S.P., M.M.N. and R.I.J. contributed single-site recurrence analysis. For the discovery of structural-variant drivers, the following authors contributed: J.A.W., O.S., Y.L., N.D.R., S.E.S., M.I. and J. Weischenfeldt contributed and curated genomic annotations; J.A.W., J.E.H., J.T., O.S., D. Craft, K.K., S.E.S., C. Stewart., C.-Z.Z., M.I., P.J.C., J. Weischenfeldt, X.Y. and R.B. contributed to the development of the structural variant recurrence analysis methods; J.A.W., O.S., K.K., J. Weischenfeldt and R.B. implemented structural variant recurrence analyses; and J.A.W., O.S., J. Busanovich, N.S., P.B., J. Weischenfeldt and R.B. integrated structural variant recurrence results with expression, chromatin organization and functional data. For point mutations candidate vetting and filtering, the following authors contributed: E.R., F.A., H.H., J. Kim, L.F., M.M.N. and T.M. contributed individual candidate filters; and A.L., C. Hong, C.W., E.K., E.M.L., E.R., F.A., G.G., G.T., H.H., H.M.U., J. Kim, J.M.H., J.S.P., K.D., L.F., L. Sieverling, M.M.N., M.S.L., N.S.-A., R.I.J., R.J. and T.M. performed candidate vetting. For candidate-based analysis, the following authors contributed: E.R., F.A., G.G., H.H., H.N., I.M., J.M.H., J.R., J.S.P., Keunchil Park, M.M.N. and M.P.H. contributed candidate-based analysis; A.G.-P., A.H., A.L., C. Hermann, D. Chakravarty, D. Tamborero, E.K., E.R., F.A., G.G., G.T., H.H., I.M., J.C.-F., J.R., J.S.P., K.I., Keunchil Park, L.M., L.D.S., L.U.-R., L.W., M. A. Rubin, M.B.G., M.M.N., M.S.L., N.S.-A., N.L.-B., O.P., R.I.J., R. Sabarinathan, S.K. and Y. Kim contributed results interpretation; A. Kahles., J.S.P., K.A.B., K.-V.L., M.M.N., N.A.F., S.B.A., T.A.J. and T.T. contributed expression profiling (extended GENCODE set); A. Kahraman, D. Chakravarty, J.R., J.S.P., K.I., L.W., M.A. Rubin, M.M.N., M.S.L., S.B.A. and T.M. contributed mutation-to-expression correlation analysis; A. Kahraman, D. Chakravarty, J.R., L.W., M.A. Rubin and N.S.-A. contributed network or pathway analysis; and R. Sabarinathan, C. Shen, C. Sander

and J.S.P. contributed structural RNA analysis. For power analysis and driver mutations at known cancer genes, the following authors contributed: E.R. analysed SNV detection and driver discovery power; Z.L. evaluated sensitivity of methods; F.A. and I.M. contributed mutational excess analysis; M.M.N. integrated additional evidence; and O.S. analysed structural variant detection and driver discovery power. The following authors contributed leadership and organizational work: for point mutations, E.R., G.G. and J.S.P. contributed working group leadership; A.G.-P., B.J.R., D.A.W., E.K., E.R., G.G., G.T., I.M., J.R., J.M.S., J.S.P., L.F., L.M., M.B.G., N.L.-B., O.P., P.J.C., R. Sabarinathan and S.K. contributed organization; and E.R., M.M.N., F.A., G.T., H.H., J.M.H., R.I.J., I.M., J.S.P. and G.G. wrote the manuscript. For structural variants, P.J.C. and R.B. contributed working group leadership; J.A.W., Y.L., P.J.C., J. Weischenfeldt and R.B. contributed organization; and J.A.W., O.S., P.J.C., J. Weischenfeldt and R.B. wrote the manuscript.

**Competing interests** The following authors declare that they have competing interests. P.B. receives grant funding from Novartis from an unrelated project; R.B. owns equity in Ampressa Therapeutics and receives grant funding from Novartis; G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMutSig and POLYSOLVER; B.J.R. is a consultant at and has ownership interest (including stock, patents and so on) in Medley Genomics; O.S. is currently an employee of Cedilla Therapeutics; and Y.L. is currently an employee of Seven Bridges Genomics.

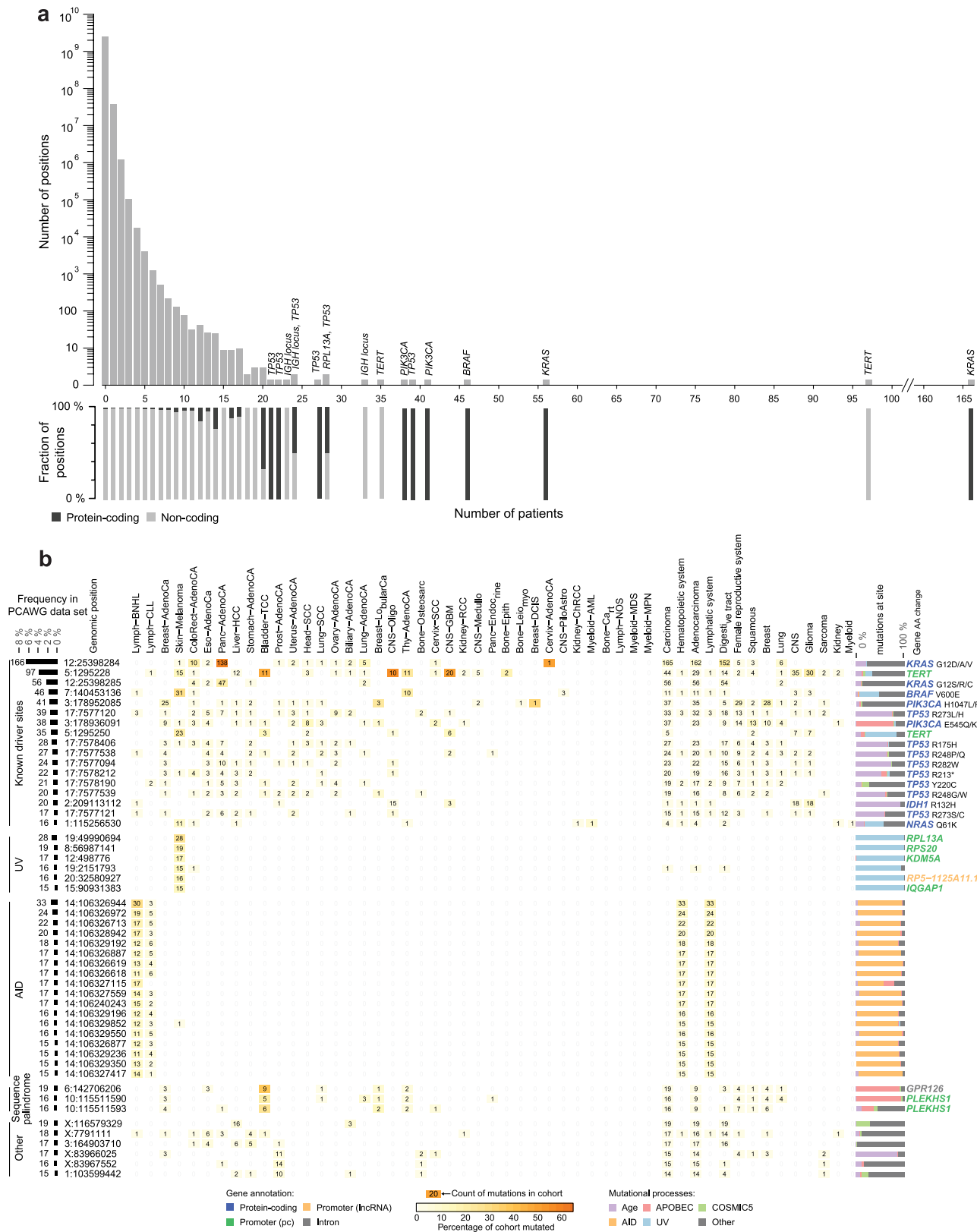
#### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1965-x>.

**Correspondence and requests for materials** should be addressed to J.Weischenfeldt, R.B., I.M., J.S.P. or G.G.

**Peer review information** *Nature* thanks Don Conrad, Fran Supek and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

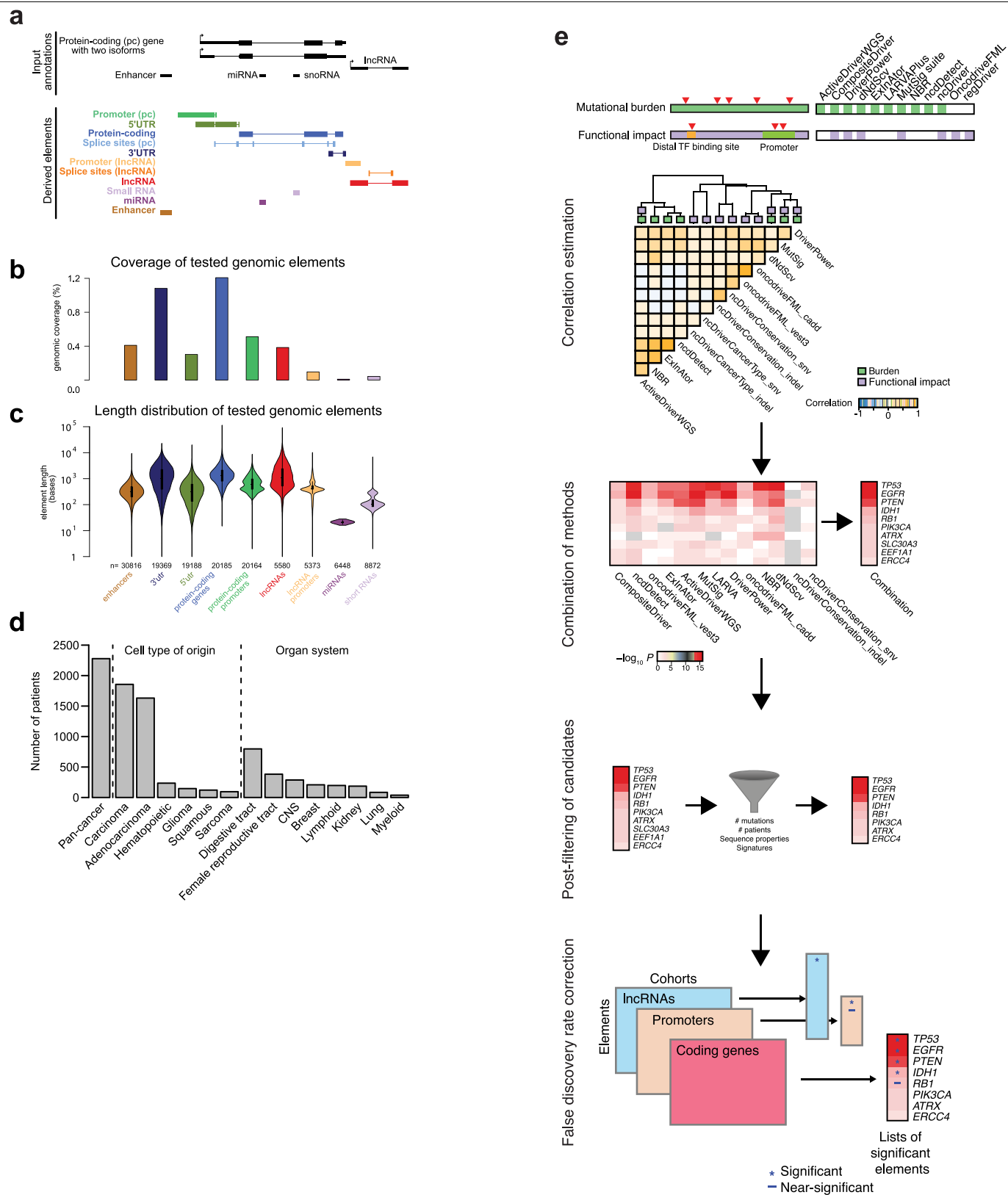
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Mutational hotspots in additional tumour types. a**, Bar plot of number of positions (y-axis) mutated in  $n$  patients (x-axis). The stacked bar charts under the bar plot show the proportion of protein-coding (dark grey) and non-coding (light grey) positions. **b**, Distribution of SNVs in top 50 single-site hotspots across all analysed individual cohorts and meta-cohorts. Hotspots are grouped as known drivers or induced by mutational processes.

The table (middle) shows the frequency of mutations across the PCAWG cohorts. Stacked bar chart (right) shows the contribution of mutational processes to the hotspot mutations (Methods). Gene names are given when hotspots overlap with functional elements (colour-coded), with amino acid alterations for protein-coding genes.

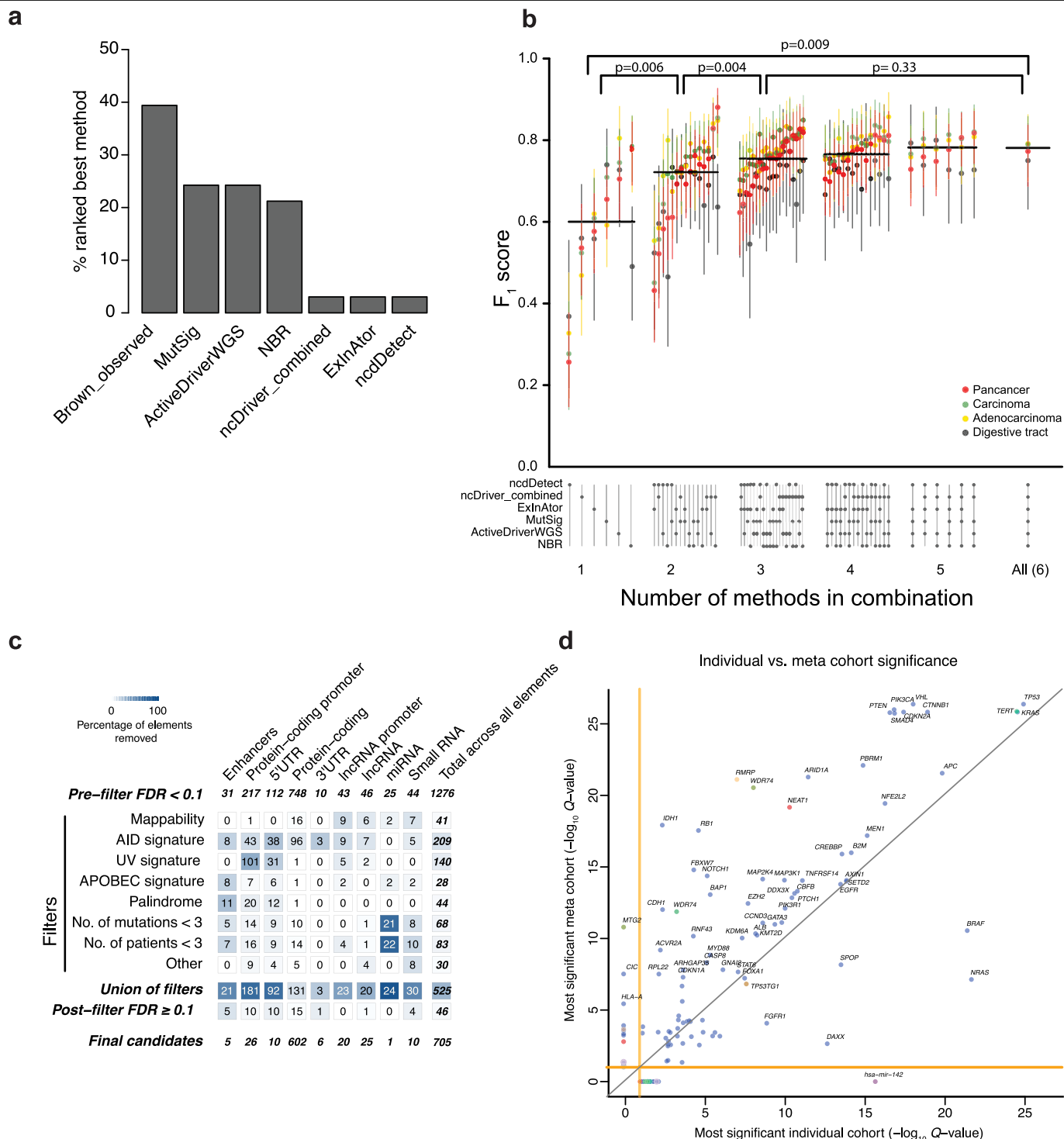




**Extended Data Fig. 2** | See next page for caption.

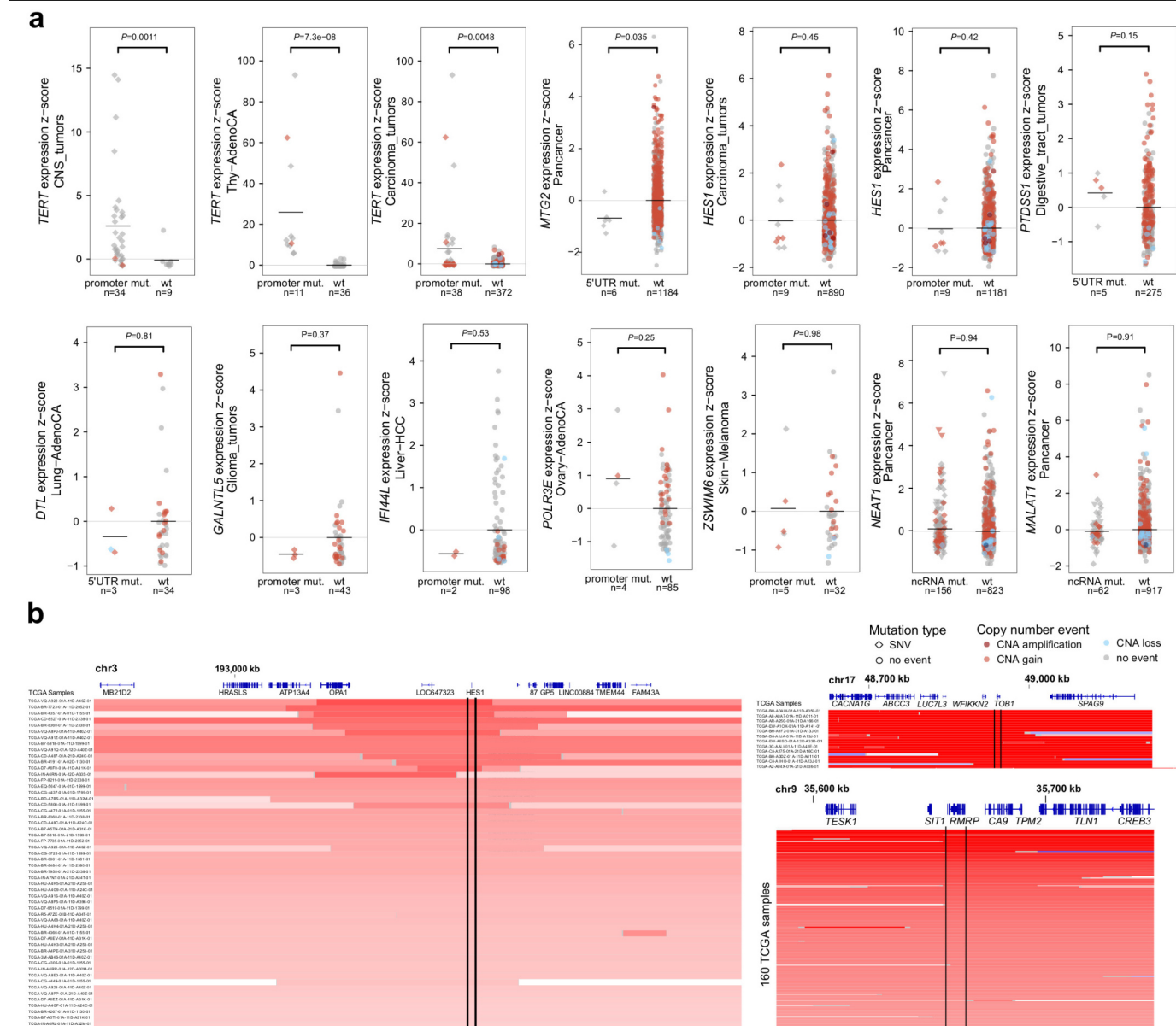
**Extended Data Fig. 2 | Element-based driver discovery and combination of *P* values.** **a**, Schematic describing definition of types of functional element (Methods). Functional elements (black) are defined on the basis of transcript annotations from various databases. Elements arising from multiple transcripts with the same gene identity are collapsed, as seen here for the protein-coding isoforms. Promoter elements are defined as 200 bases upstream and downstream of the transcription start sites of the transcripts of a gene (green). Splice site elements extend 6 and 20 bases from the 3' and 5' exonic ends into intronic regions, respectively (light blue). Regions overlapping protein-coding bases and protein-coding splice sites are subtracted from other regions. **b**, Percentage of genomic coverage for each element type. **c**, Distribution of element lengths for each element type. Thick lines indicate interquartile ranges and short horizontal bars indicate the medians. **d**, Organization of meta-cohorts defined by tissue of origin and organ system. Pan-cancer contains all cancers, excluding Skin–Melanoma and

lymphoid malignancies. **e**, Combination workflow: overview of methods of driver discovery and their lines of evidence to evaluate candidate gene drivers. Methods using each feature are marked with a box in the appropriate track. Heat map displaying Spearman's correlation of *P* values across the different driver-discovery algorithms based on simulated (null model) mutational data. Dendrogram illustrates the relatedness of method *P* values, and algorithm approaches are marked by coloured boxes on dendrogram leaves. Next, *P* values are combined with Brown's method on the basis of the calculated correlation structure. Individual method (left) and integrated (right) log-transformed *P* values are shown in a heat map (grey, missing data). Post-filtering used several criteria to identify likely suspicious candidates. Significant driver candidates were identified after controlling for multiple hypothesis testing based on an FDR *Q* value threshold of 0.1 (blue asterisk). Candidates with *Q* values below 0.25 (blue dash) were also considered of interest.



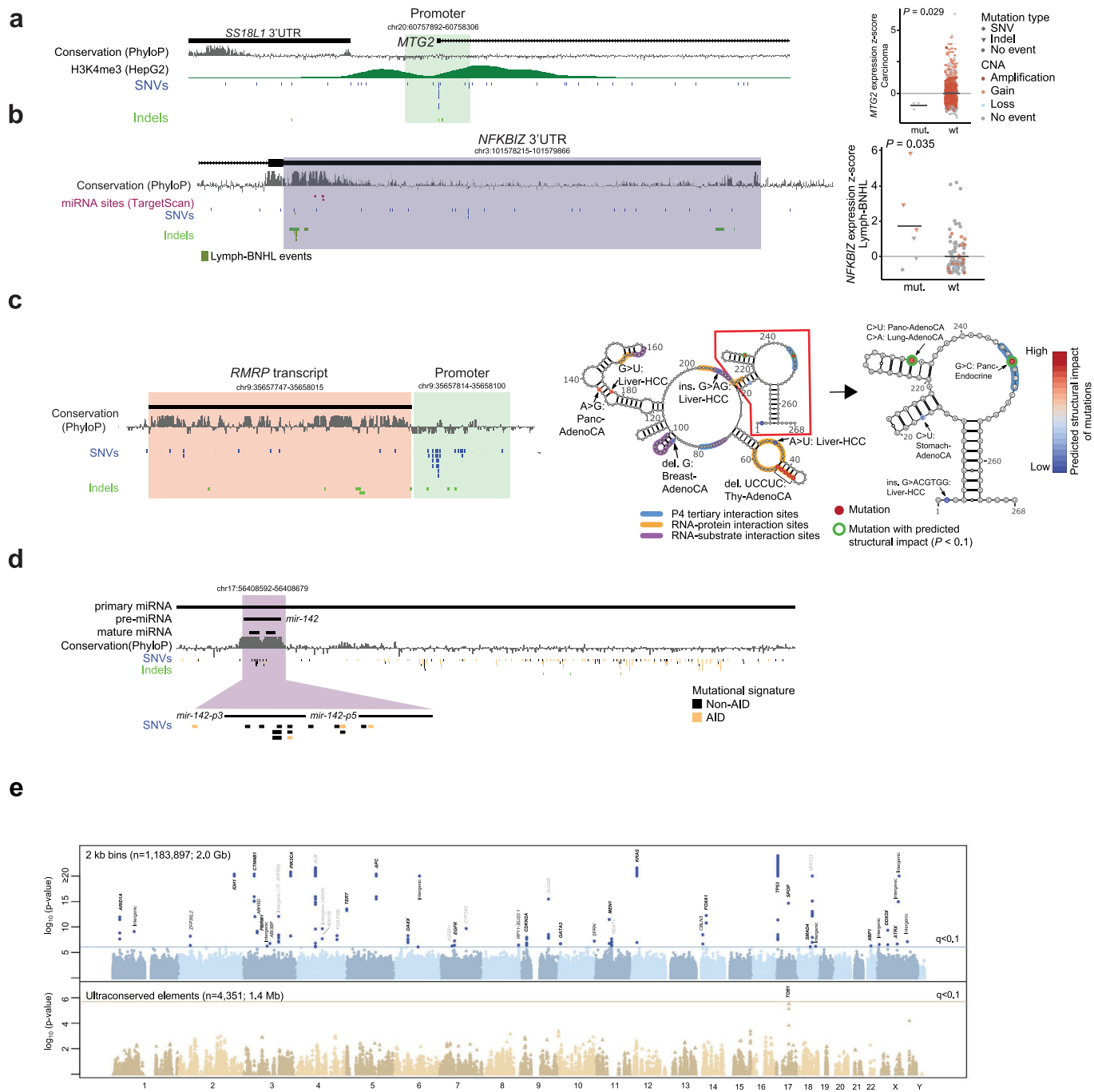
**Extended Data Fig. 3 | Sensitivity of driver-discovery methods and filter statistics.** **a**, Percentage of coding-driver discovery runs (with stable  $F_1$  score,  $n = 33$ ), across all cohorts, in which the method had the highest  $F_1$  score (Methods). **b**,  $F_1$  score of different methods of driver discovery, and different combinations evaluated in the four largest cohorts (pan-cancer ( $n = 2,278$ ), carcinoma ( $n = 1,856$ ), adenocarcinoma ( $n = 1,631$ ) and digestive tract ( $n = 797$ )). Only methods that used the same algorithm to call coding and non-coding drivers were evaluated. Vertical lines indicate 95% confidence intervals. Horizontal black lines mark the median in each group.  $P$  values were calculated with the two-sided non-parametric Mann-Whitney  $U$  test. **c**, On top, the initial number of hits identified as recurrently mutated for each element type. The element types mature miRNA ( $n = 2$  before filtering) and miRNA promoters

( $n = 16$  before filtering) were omitted from the table. The heat map shows the number of hits filtered at each step in the sequential application of filters and post-filtering re-application of the FDR correction. Background colours indicate the corresponding percentage of input element removed. The final numbers of hits (including those that were later filtered by the comprehensive vetting procedures) are indicated below the heat map. **d**, Sensitivity versus specificity in individual cohorts versus meta-cohorts for candidate drivers:  $Q$  values for the most significant individual cohort ( $x$  axis) versus meta cohort ( $y$  axis) are shown. Driver elements are coloured by their element type.  $Q$  values derived from combination of  $P$  values from individual driver-discovery methods (Methods).



**Extended Data Fig. 4 | Mutation-to-expression correlation and focal copy-number alterations. a**, Expression is compared between mutated and non-mutated samples. For each element, the z score of the expression values for mutated and wild type in the significant cohort is plotted. For copy number, CNA amplification indicates CNA > 10; CNA gain indicates CNA ≥ 3; CNA loss indicates CNA ≤ 1; and no events indicates CNA < 3 and CNA > 1. If a patient is mutated with multiple types of point mutation, indels are indicated over SNVs.

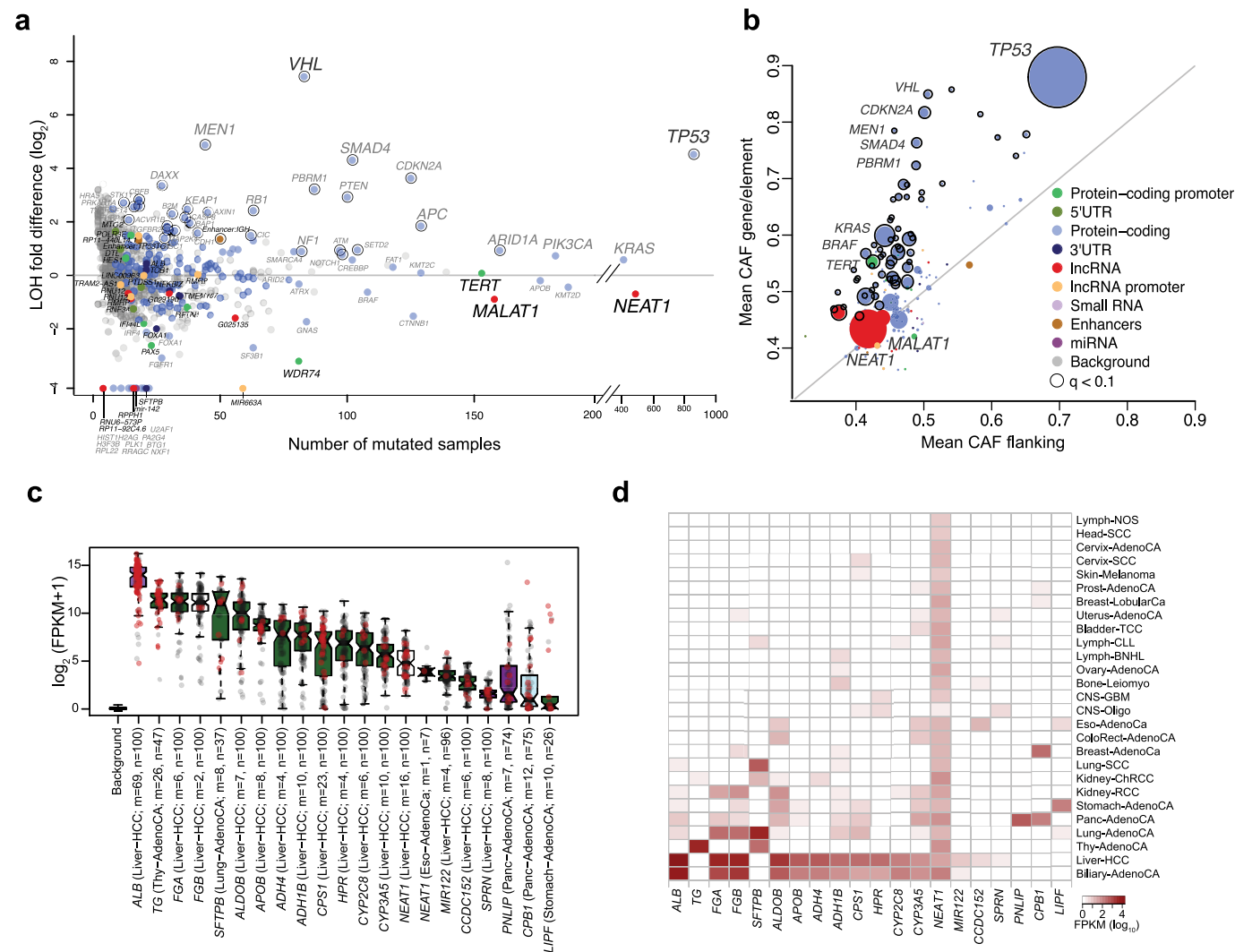
For *TERT*, only samples powered to call mutation status were used. *P* values are based on a two-sided Wilcoxon rank-sum test. Bars indicate means. **b**, Copy-number profiles of 55 of 441 stomach adenocarcinomas from TCGA show copy-number gains around *HES1*, *TOB1* and its gene neighbour *WFIKN2* are focally amplified in cancer (172 of 10,844 total samples from 33 cancer types are shown). *RMRP* focal amplifications in TCGA cancers (160 of 10,844 total tumours shown).



**Extended Data Fig. 5 | Non-coding driver candidates. a**, *MTG2* promoter locus (left) and associated gene-expression changes in carcinoma tumours (right). Expression of *MTG2* in mutated ( $n = 3$ ) versus the carcinoma meta-cohort wild-type cases ( $n = 896$ ). Two-sided Wilcoxon rank-sum test. Bars represent means. **b**, Genomic locus of *NFKB1Z* 3'UTR (left) and associated gene-expression changes in Lymph-BNHL (right). Expression of *NFKB1Z* in mutated ( $n = 6$ ) versus wild-type cases ( $n = 98$ ). Test and bars as in **b**. **c**, Genomic locus of the *RMRP* transcript and promoter region (left). *RMRP* is an RNA component of the endoribonuclease RNase MRP, the function of which depends on its RNA

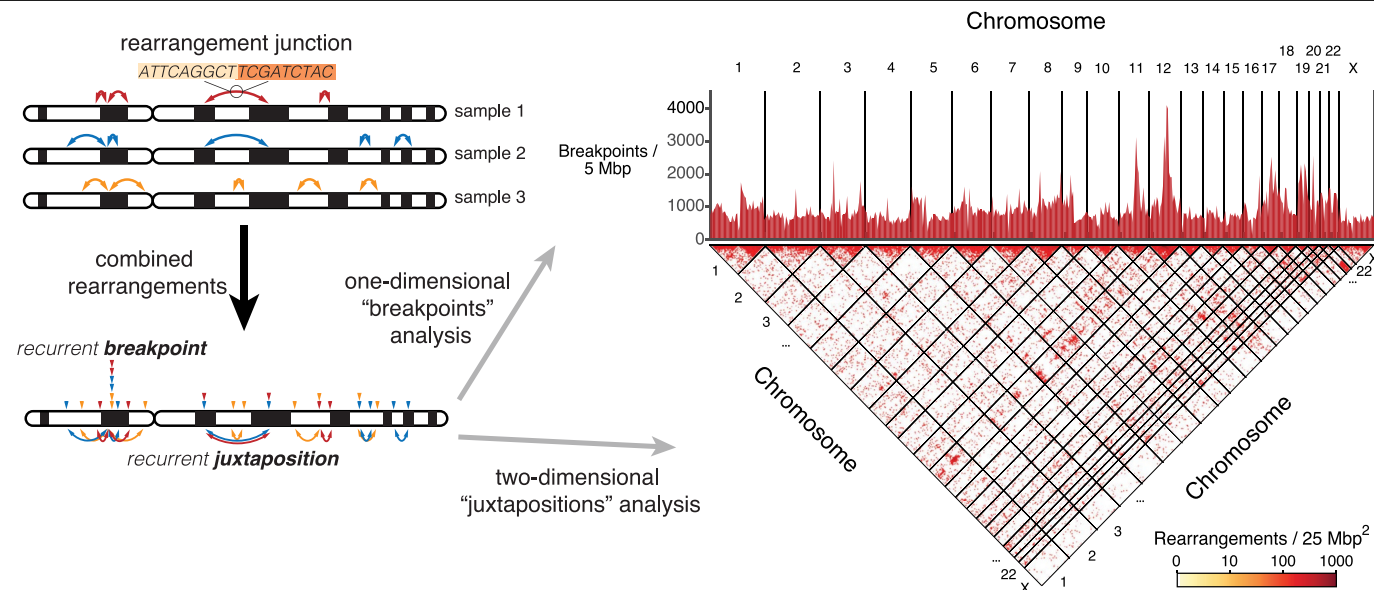
secondary and tertiary structure. The RNA secondary structure, tertiary structure interactions, protein and substrate interactions, and mutations with their predicted structural effect (right) of *RMRP*; lymphoma and melanoma mutations are excluded. **d**, *MIR142* locus and mutations in patients with lymphoma with the AID signature annotation. **e**, Manhattan-style plot showing significance of mutation recurrence enrichment for genomic bins (top) and ultraconserved elements (bottom) across cohorts (Methods; Supplementary Table 9).





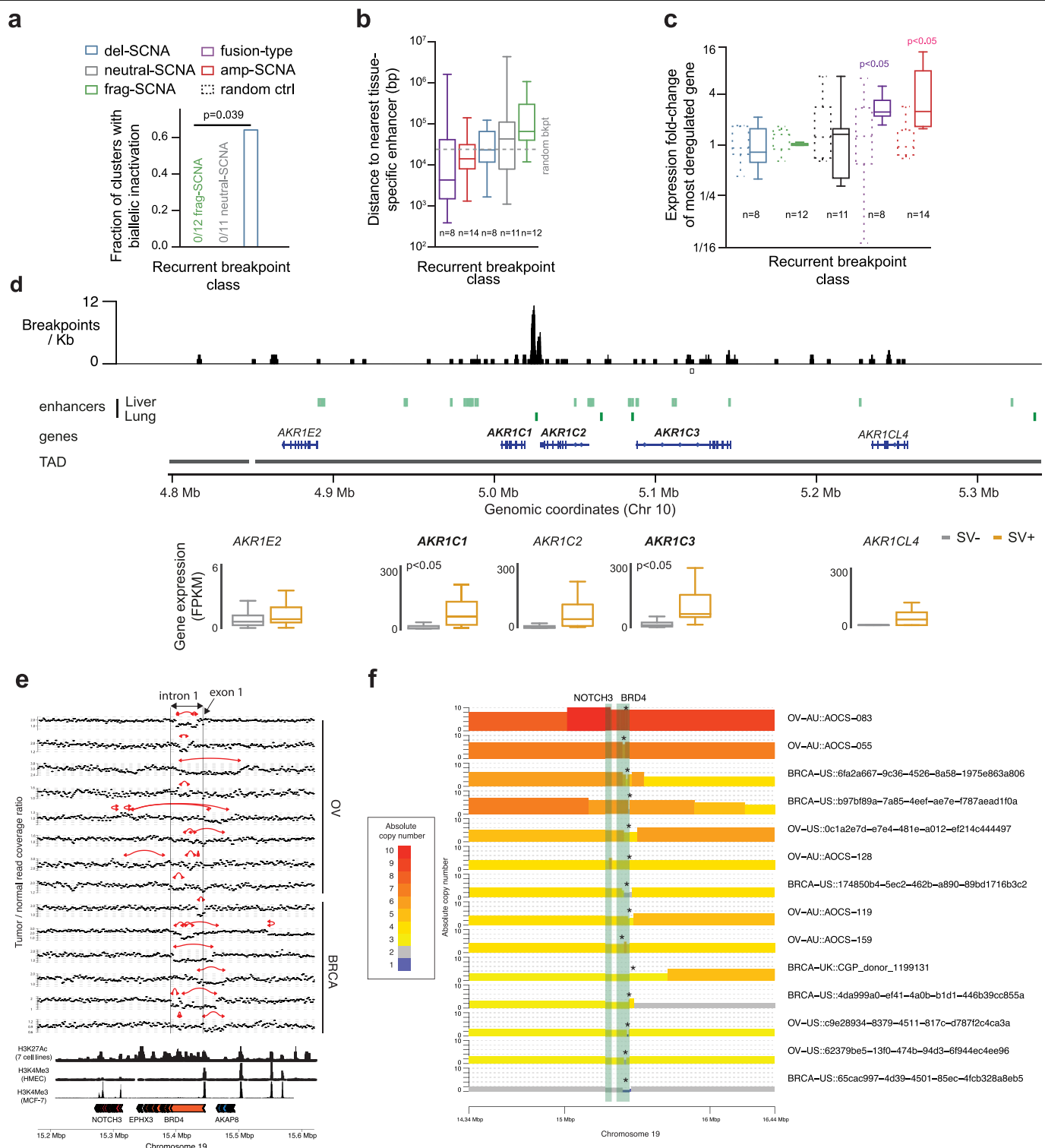
**Extended Data Fig. 6 | A transcriptional process creates passenger mutations in highly expressed, tissue-specific genes. a**, Relative rate of loss-of-heterozygosity (LOH) compared between mutated and wild-type samples for all significant elements, coloured by element type and highlighting significant LOH enrichments with an outside black circle (Fisher's exact test, one-sided;  $Q < 0.1$ ). **b**, Average cancer allelic fraction (CAF) compared between each significant genomic element and the corresponding flanking regions ( $\pm 2$  kb and introns; overlapping coding exons were excluded). The size of the

points represents the number of mutated samples for each particular element. Genes with significantly higher CAFs ( $t$ -test, one-sided;  $Q < 0.1$ ) are highlighted with an outside black circle. **c**, mRNA expression of genes enriched in 2-5-bp indels in their respective tissues. Boxes show the interquartile range and median. The first box contains background gene-expression levels. Red and grey dots correspond to samples with ( $m$ ) and without ( $n - m$ ) indels in the corresponding gene. **d**, Heat map showing the levels of expression across types of cancer for the genes enriched in 2-5-bp indels.



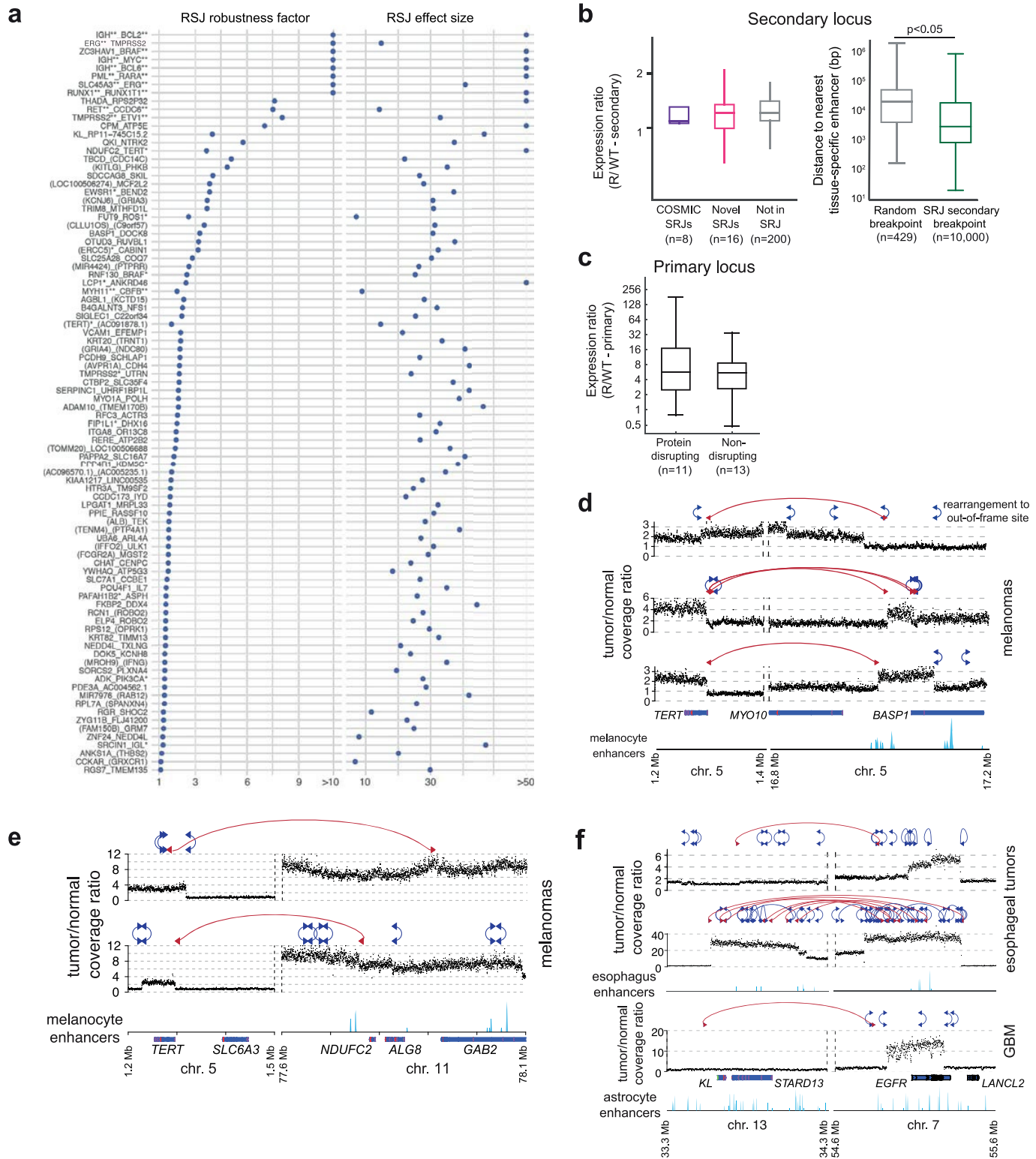
**Extended Data Fig. 7 | Overview of structural-variant analysis.** Schematic indicating analysis approach. Left, rearrangements and rearrangement junctions in three hypothetical genomes (top) and the two analysis approaches (bottom): the 1D analysis for recurrent breakpoints and the 2D analysis for

recurrent juxtapositions between pairs of loci. Right, the 1D density of breakpoints genome-wide (top) and 2D density of juxtapositions (bottom) across 2,693 cancer genomes (Methods).



**Extended Data Fig. 8 | Gene-expression effects of SRBs. a**, Fraction of recurrent breakpoint loci associated with biallelic inactivation of a known tumour suppressor gene (frag-SCNA, 0/12; neutral-SCNA, 0/14; del-SCNA, 5/8; Fisher's exact test). **b**, Distance in bp to the nearest tissue-specific enhancer for each breakpoint class. Dashed grey line represents 1,000 randomly selected breakpoints from the same tumour samples. All box plots show the interquartile range, median and 95% confidence interval. **c**, Expression fold change for the gene with the most-altered expression within 1 Mb of the cluster centroid in samples with, compared to samples without, a breakpoint at the cluster locus. Random controls (in dashed boxes) represent 1,000 randomly selected breakpoints. *P* values are from two-sided *t*-tests (Methods). **d**, Breakpoint density near AKRIC genes (top), locations of enhancers (middle)

and expression of local genes (bottom;  $n=7$  SV+ tumours,  $n=41$  SV- lung squamous cell tumours; two-sided *t*-test) in samples with and without local rearrangements. **e**, Ratio of tumour-to-normal read coverage across six breast tumours and eight ovarian tumours with focal *BRD4* exon 1 and intron 1 deletions. Red lines indicate rearrangements. **f**, Amplification structure (absolute copy number, *y* axis) of the *BRD4* and *NOTCH3* locus in breast and ovarian tumours with a *BRD4* focal deletion. In most cases, the copy-number caller identified the focal deletion. However, in some cases, the deletions were too small to be identified only using read depth. When combining read depth and rearrangement signals in **a**, there is clear evidence for focal deletions. Deletion locations are marked by an asterisk.

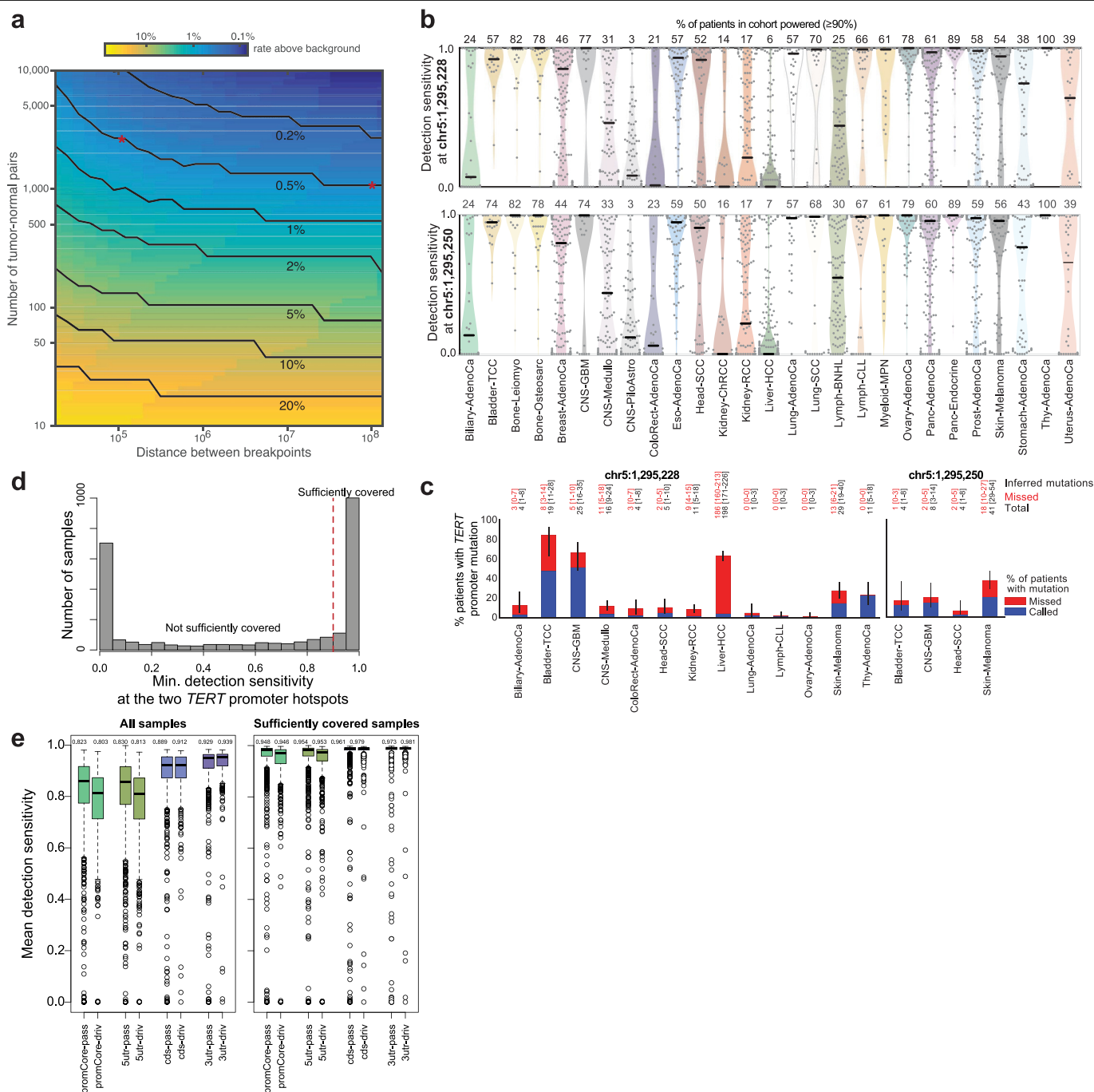


**Extended Data Fig. 9** | See next page for caption.

**Extended Data Fig. 9 | Gene-expression effects of SRJs.** **a**, Assessment of SRJ robustness against unaccounted for mechanistic and technical confounders. Left, a robustness factor, defined as the ratio between the background probability value that would lower the *P* value of an SRJ below the genome-wide *P*-value threshold and the estimator for the background probability from our 2D model. Higher robustness values represent lower susceptibility to unaccounted variations in the background model. The top 48 SRJs have a robustness factor greater than 2, which suggests that these SRJs would remain significant even if the true background rate was twice as high as our model estimates. Right, the effect size is calculated as the difference in observed and estimated number of SRJs in units of standard deviation (assuming binomial distribution of structural variant count per 2D genomic region). Most SRJs are well above ten standard deviations of the predicted value. **b**, Characteristics of SRJ secondary loci. Left, fold expression enrichment of the most highly overexpressed gene in the secondary locus in cancer samples with these

fusions relative to cancers of the same histology without the fusion. Right, the distance from the SRJ secondary locus (green) to the nearest enhancer is significantly smaller ( $P < 0.05$ ; two-sided *t*-test) compared to randomly selected breakpoints (grey). **c**, Fold expression enrichment of the most highly overexpressed gene in the primary locus, for fusions that disrupt protein-coding sequences and fusions that do not. All box plots show the interquartile range, median and 95% confidence interval. **d**, Rearrangements between the *TERT* promoter and the *BASPI* and *MYO10* locus result in focal amplification of *TERT* and relocation of distal enhancers to *TERT*. **e**, *TERT-NDUFC2* fusion in two melanoma samples connecting *TERT* with an enhancer-rich region next to *NDUFC2*. Both samples also have focal amplifications of *TERT*. **f**, Recurrent translocation between *EGFR* in chromosome 7 and the *KL* and *STARD13* locus on chromosome 13. In all three samples, the rearrangement contributed to the amplification of *EGFR*.





**Extended Data Fig. 10 | A lack of detection power in specific elements.**

**a**, Number of tumour-normal pairs needed to detect fusions with 90% power as a function of the span of the fusion and the rate above background at which it recurs. The red asterisks indicate the numbers of samples required to detect 100-kb and 100-Mb fusions that recur at 0.5% above their background rates.

**b**, Distribution of *TERT* promoter hotspot (top, chromosome 5:1,295,228; bottom, chromosome 5:1,295,250; hg19) detection sensitivity for each patient, by cohort. Grey dots indicate values for individual patients inside estimated distribution (areas coloured by cohort). Horizontal black bars mark the medians. Numbers above distributions indicate the percentage of patients powered (detection sensitivity  $\geq 90\%$ ) in each cohort. Cohort sizes as in Fig. 4a.

**c**, Percentage of patients with observed (blue) and inferred missed (red) mutations at the chromosome 5:1,295,228 and chromosome 5:1,295,250 *TERT*

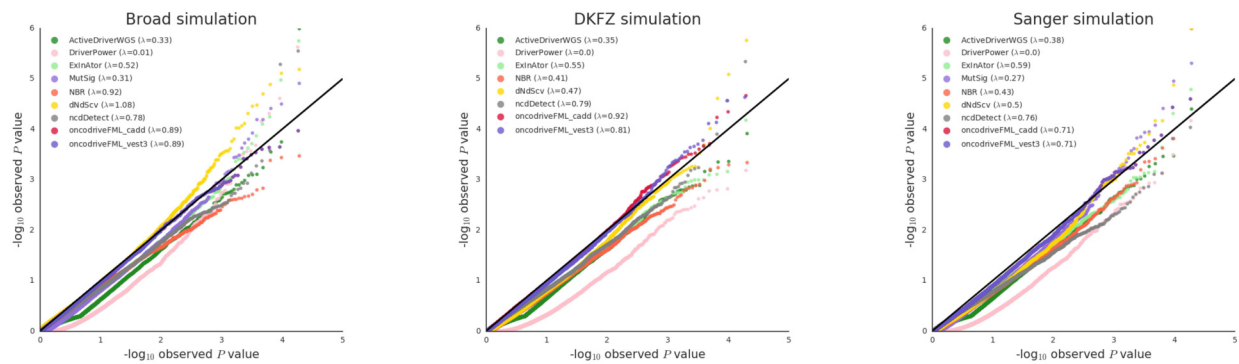
promoter hotspot sites. Error bars indicate 95% Poisson confidence interval.

Numbers above bars show the total inferred number of *TERT* promoter mutations for each site in this cohort. Red numbers indicate the absolute number of inferred missed mutations (owing to a lack of read coverage).

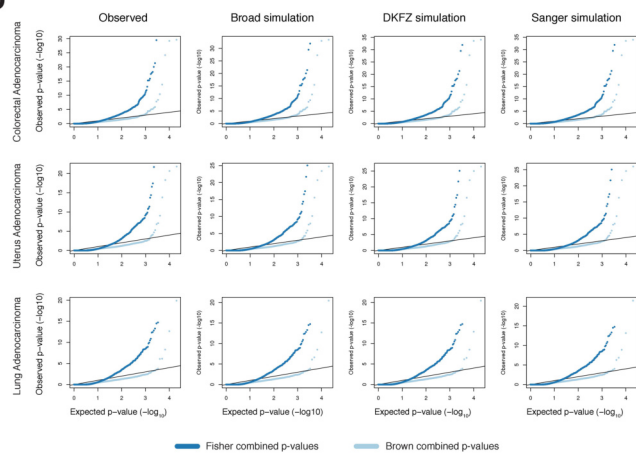
Cohort sizes as in Fig. 4a. **d**, Detection sensitivity for the two *TERT* promoter hotspots across all samples showing the variation in powered samples. Red vertical line ( $x = 0.9$ ) indicates cutoff for 'sufficiently powered samples'.

**e**, Mean detection sensitivity in 1,000 randomly selected putative passengers (pass) and 603 cancer genes (driv) across element types: promoters, 5' UTRs, CDS and 3' UTRs. The left panel shows the results for all samples and the right panel corresponds to the set of samples with high sensitivity at *TERT* hotspots. Boxes show the interquartile range and median; outliers are shown as circles. Weighted sensitivity means are shown at the top of the box plot.

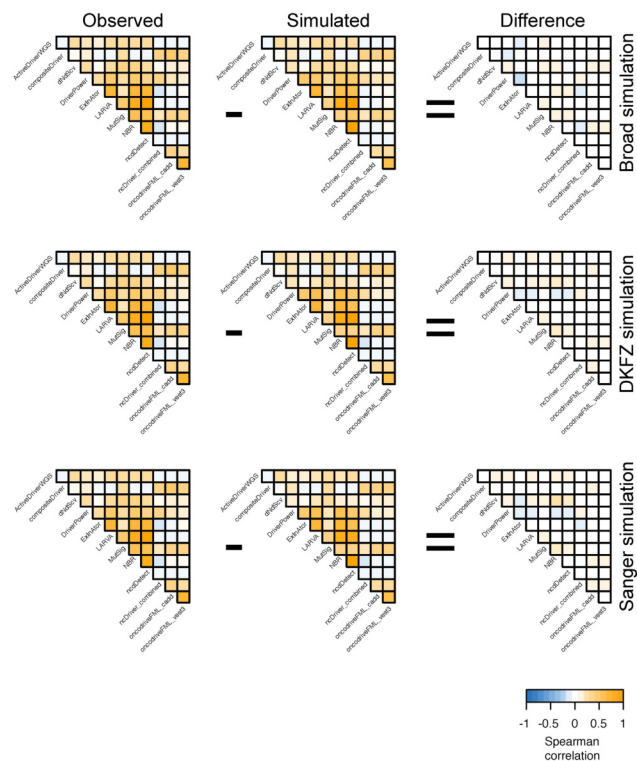
a



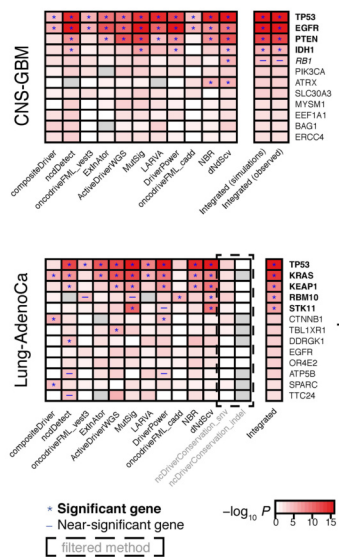
b



c



d



Extended Data Fig. 11 | See next page for caption.

**Extended Data Fig. 11 | P value combination details.** **a**, Quantile–quantile plots of *P* values reported by various driver-detection algorithms on the three simulated datasets (Broad, DKFZ and Sanger; shown for coding regions ( $n = 20,172$ ) in the meta-carcinoma cohort; see Methods for details for the statistical background model or test of each algorithm) showed no major enrichment of mutations above the background rate. Results generally followed the expected null (uniform) distribution, and the *P* values reported on simulated data were subsequently used to assess the covariance of method results. **b**, Quantile–quantile plots of integrated *P* values using the Brown and Fisher methods for combining *P* values across the results from different driver-detection algorithms were generated for a few representative tumour cohorts (shown here for coding regions). Brown combined *P* values (light blue) generally followed the null distribution as expected, whereas Fisher combined *P* values were significantly inflated (dark blue), confirming that dependencies existed between the results reported by the various driver-detection algorithms. To simplify the integration procedure, we calculated covariances using *P* values from the observed data instead of simulated data and found that

the integrated results based on the observed covariances (first column of plots) were essentially the same as the results obtained using the simulated covariances (second, third, and fourth columns of plots). **c**, Triangular heat maps showing the Spearman correlations of *P* values among the various driver-detection methods in observed versus simulated data (coding regions ( $n = 20,172$ ), colorectal adenocarcinoma cohort) are highly similar. Differences in the observed and simulated correlation values (shown in the heat maps on the far right) were minimal, and thus the final integration of *P* values across methods was performed using covariances estimated on observed data. **d**, Brown combined *P* values based on observed and simulated covariance estimations (shown on the right, top heat map, for coding regions in glioblastoma) did not differ noticeably. In cases in which individual methods reported results that yielded substantially fewer hits than the median across all methods (bottom heat map, methods in light grey with results in dashed box), removing the methods from the integration did not affect the number of significant genes identified (right column of results in bottom heat map, shown for coding regions in lung adenocarcinoma). Number of coding regions as in **c**.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

Data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at <https://www.overture.bio>. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to [contact@overture.bio](mailto:contact@overture.bio).

#### Data analysis

The workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: <https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files>. Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACESeq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBBA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15

The method for combining p-values is available from [https://github.com/broadinstitute/getzlab-PCAWG-pvalue\\_combination/](https://github.com/broadinstitute/getzlab-PCAWG-pvalue_combination/). Power calculations are available from [https://github.com/broadinstitute/getzlab-PCAWG-power\\_calculations](https://github.com/broadinstitute/getzlab-PCAWG-power_calculations). The method for identifying significantly recurrent breakpoints by controlling for covariates is available at: <https://github.com/mskilab/fish.hook>. The method for permuting rearrangement breakpoint pairs to identify covariates affecting rearrangement partner selection is available as the "swap" module of: <https://github.com/walaj/ginseng>. The method for identify significantly recurrent rearrangements is available as the "2D" method at: <https://github.com/ofershapira/SVsig>. The method for identifying cis-expression consequences of SVs, CESAM, is available at: <https://bitbucket.org/weischen/cesam>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

WGS somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. For analyses specific to this study, we generated sample subsets as described in the methods.
Data exclusions	After quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine).
Replication	In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement.
Randomization	No randomisation was performed.
Blinding	No blinding was undertaken.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging



## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.
Recruitment	Patients were recruited by the participating centres following local protocols.
Ethics oversight	The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Patterns of somatic structural variation in human cancer genomes

<https://doi.org/10.1038/s41586-019-1913-9>

Received: 22 September 2017

Accepted: 18 November 2019

Published online: 5 February 2020

Open access

Yilong Li<sup>1,2,14</sup>, Nicola D. Roberts<sup>1,14</sup>, Jeremiah A. Wala<sup>3,4,5,14</sup>, Ofer Shapira<sup>3,4,5,14</sup>, Steven E. Schumacher<sup>3,4,5</sup>, Kiran Kumar<sup>3,4,5</sup>, Ekta Khurana<sup>6</sup>, Sebastian Waszak<sup>7</sup>, Jan O. Korbel<sup>7</sup>, James E. Haber<sup>8</sup>, Marcin Imielinski<sup>9</sup>, PCAWG Structural Variation Working Group<sup>10</sup>, Joachim Weischenfeldt<sup>11\*</sup>, Rameen Beroukhi<sup>3,4,5\*</sup>, Peter J. Campbell<sup>1,12\*</sup> & PCAWG Consortium<sup>13</sup>

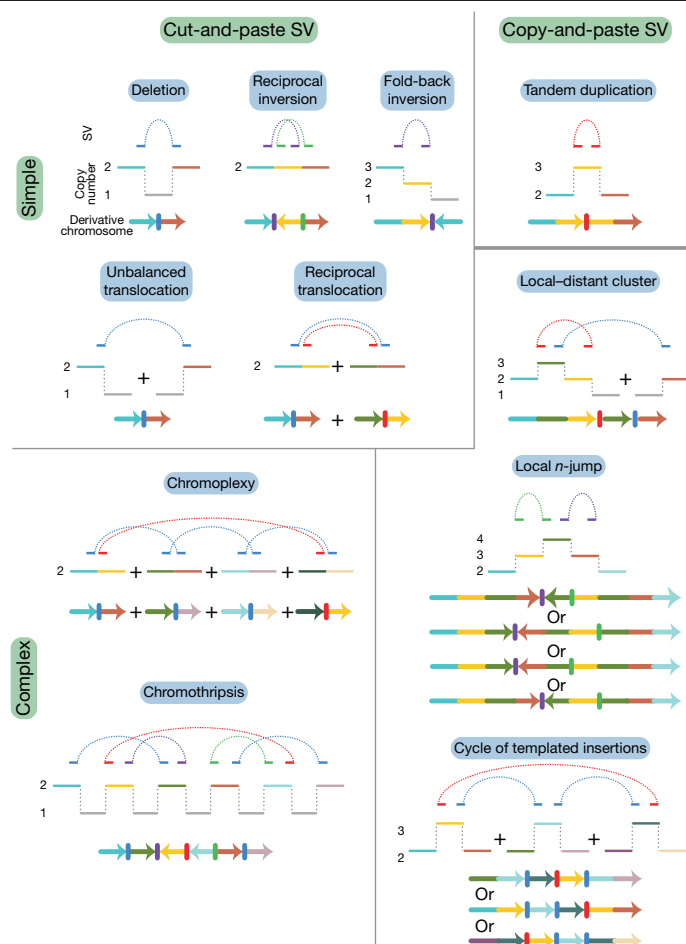
A key mutational process in cancer is structural variation, in which rearrangements delete, amplify or reorder genomic segments that range in size from kilobases to whole chromosomes<sup>1–7</sup>. Here we develop methods to group, classify and describe somatic structural variants, using data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), which aggregated whole-genome sequencing data from 2,658 cancers across 38 tumour types<sup>8</sup>. Sixteen signatures of structural variation emerged. Deletions have a multimodal size distribution, assort unevenly across tumour types and patients, are enriched in late-replicating regions and correlate with inversions. Tandem duplications also have a multimodal size distribution, but are enriched in early-replicating regions—as are unbalanced translocations. Replication-based mechanisms of rearrangement generate varied chromosomal structures with low-level copy-number gains and frequent inverted rearrangements. One prominent structure consists of 2–7 templates copied from distinct regions of the genome strung together within one locus. Such cycles of templated insertions correlate with tandem duplications, and—in liver cancer—frequently activate the telomerase gene *TERT*. A wide variety of rearrangement processes are active in cancer, which generate complex configurations of the genome upon which selection can act.

Mutations that arise in somatic cells are the driving force of cancer development. Structural variation—in which genomic rearrangement acts to amplify, delete or reorder chromosomal material at scales that range from single genes to entire chromosomes—is an especially important class of somatic mutation. Previous analyses of both cancer and germline genomes have enabled the description of several distinctive patterns of structural variants<sup>1–7</sup>, and hypotheses about the underlying basis of several of these patterns have been proposed on the basis of their clustering, orientation and associated copy-number changes. Hypothesis-driven *in vitro* studies are now beginning to reveal some of the mechanistic processes that generate these structures<sup>9–13</sup>, and generate further predictions that can be assessed in the genomic data. However, the landscape of structural variation in human cancer remains incompletely mapped and there are many complex structures that elude formal description.

The PCAWG Consortium aggregated whole-genome sequencing data from 2,658 cancers across 38 tumour types, generated by the ICGC

and TCGA projects. These sequencing data were aligned to the human genome (reference build hs37d5) and analysed with standardized, high-accuracy pipelines to call somatic and germline variants of all classes<sup>8</sup>. Here, we analyse the patterns and signatures of structural variants across the PCAWG data. We propose a working classification scheme that encompasses known and newly identified classes of structural variants. We develop methods for annotating the observed structural variants in a given cancer genome, identifying a class of replication-based rearrangement processes that generate clusters of several structural variants. We explore the size, activity and genome-wide distribution of classifiable structural variant types across the cohort, using signature analysis to define how they correlate within patients. Other papers produced by PCAWG address complementary aspects of structural variants, including inference of positive selection acting on recurrently rearranged regions of the genome<sup>14</sup>, how structural variants affect the transcriptome<sup>15</sup> and chromosome topology<sup>16</sup>, patterns of somatic retrotransposition<sup>17</sup> and distribution of chromothripsis across cancer types<sup>18</sup>.

<sup>1</sup>Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. <sup>2</sup>Totient Inc, Cambridge, MA, USA. <sup>3</sup>The Broad Institute of Harvard and MIT, Cambridge, MA, USA. <sup>4</sup>Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA, USA. <sup>5</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>6</sup>Weill Cornell Medical College, New York, NY, USA. <sup>7</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. <sup>8</sup>Department of Molecular Biology, Rosenthal Basic Medical Sciences Research Center, Brandeis University, Waltham, MA, USA. <sup>9</sup>New York Genome Center, New York, NY, USA. <sup>10</sup>A list of members and their affiliations appears at the end of the paper. <sup>11</sup>Biotech Research & Innovation Centre (BRIC), The Finsen Laboratory, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. <sup>12</sup>Department of Haematology, University of Cambridge, Cambridge, UK. <sup>13</sup>A list of members and their affiliations appears in the Supplementary Information. <sup>14</sup>These authors contributed equally: Yilong Li, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira. \*e-mail: joachim.weischenfeldt@bric.ku.dk; rameen\_beroukhi@dfci.harvard.edu; pc8@sanger.ac.uk



**Fig. 1 | Classification of structural variants in cancer genomes.** Schematics of major structural-variant (SV) classes, grouped according to whether they are simple or complex and arise through cut-and-paste or copy-and-paste processes. Each schematic comprises three parts. The top segment shows dotted arcs for each rearrangement junction that joins two chromosomal segments together. The middle segment shows the copy number of genomic segments that are involved. The bottom segment shows the configuration of the final derivative chromosome that results from the structural variant; the colour of the segments corresponds to the colour of that segment in the copy-number schematic. + indicates the different derivative chromosomes created for some of the classes: that is, the structural variants are not phased to a single derivative.

## Classification of structural variants

A 'structural variant' manifests as a 'junction' between two 'breakpoints' in the genome (terms in inverted commas here and below refer to those defined in the glossary in Extended Data Table 1). Generally, there will be a change in copy number across a given breakpoint if only one side of the break is rescued by a structural variant; if both sides of a double-stranded DNA break are rescued, a 'reciprocal' or 'balanced' structural variant will result, without substantial copy-number change. We sometimes observe 'clusters of structural variants' in which several breakpoints occur close together, in time or in genomic space—usually both. Such spatial and/or temporal proximity generally, but not always, implies that the structural variants within a cluster are mechanistically linked. Clusters can be 'phased' (in which case all structural variants in the cluster resolve to a single derivative chromosome) or 'unphased', in which case the structural variants are carried on different derivative chromosomes. An example of the latter is a reciprocal translocation that results in two derivative chromosomes, each with a single inter-chromosomal breakpoint junction (Fig. 1).

We recognize distinct 'classes of structural variant' from the orientation of the two segments at the junction and associated copy-number changes (Fig. 1, Supplementary Fig. 1). Some classes of structural variant (such as isochromosomes and rearrangements between extended, highly homologous sequences) are difficult to detect with short-read sequencing data; these classes are not considered further here. We propose categorizing classes of structural variant across two facets: the number of breakpoints involved (simple or complex) and by whether the patterns are likely to arise from 'cut-and-paste' or 'copy-and-paste' rearrangement processes. A cut-and-paste process generates a cluster of structural variants consistent with reshuffling or loss of extant genomic segments, and a copy-and-paste process is one in which copies of genomic 'templates' are newly replicated or synthesized and inserted during the rearrangement process. Deletions, reciprocal inversions, unbalanced translocations and reciprocal translocations are examples of simple cut-and-paste structural variants, as they can be reconstructed from the incorrect religation of chromosomal breaks. Tandem duplications are simple copy-and-paste structural variants, as they arise through the local insertion of a newly generated extra copy of a genomic template.

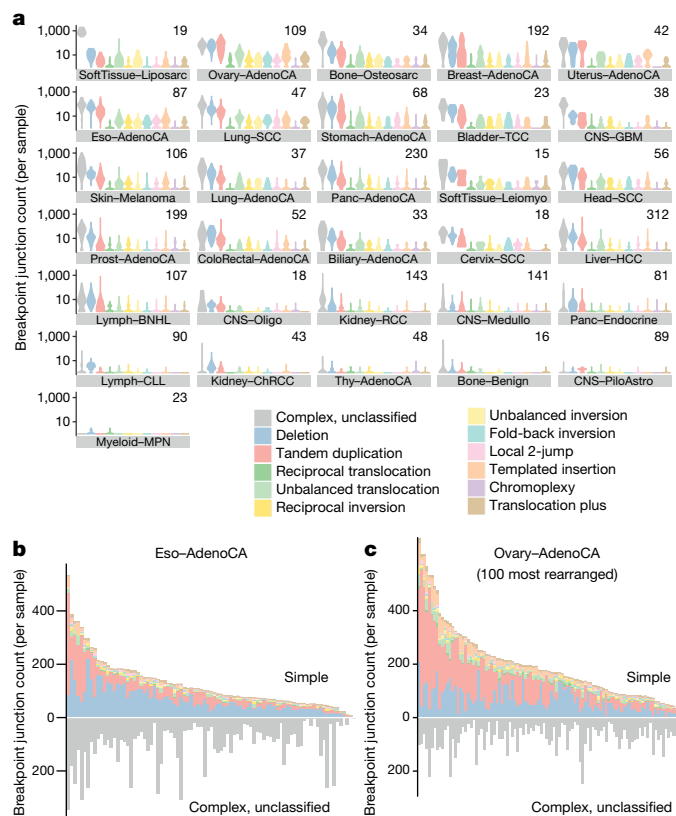
More-complex cut-and-paste processes that produce structural variants also occur in cancer. 'Breakage–fusion–bridge' events result from cycles of DNA breakage, end-to-end sister chromatid fusions, mitotic bridges and further DNA breakage. These events manifest as one or a few proximate, inverted breakpoint junctions with associated copy-number change, which we call 'fold-back inversions'<sup>1,2,19</sup> (Fig. 1). 'Chromoplexy'<sup>5,20</sup>—which is particularly frequent in prostate cancers—results from several simultaneous double-stranded DNA breaks in several chromosomes that are rejoined incorrectly, leading to balanced chains of rearrangements. 'Chromothripsis'<sup>3</sup>, in which chromosome shattering and rearrangement occur in a single catastrophic event<sup>9,21</sup>, leads to a pattern of oscillating copy-number changes and localized clustering of tens to hundreds of breakpoints<sup>22</sup>.

In the germline, more-complex copy-and-paste classes of structural variant have previously been described, which involve small duplications and triplications and are thought to arise from the stalling of the replication fork leading to template switching<sup>4,23,24</sup>. Here we describe a wide range of complex copy-and-paste types of somatic structural variant that occur in human cancers, and that are typically characterized by copy-number gains and frequent inverted rearrangements.

## Annotation of structural-variant classes

We analysed 2,559 whole cancer genomes across 38 tumour types (alongside matched germline DNA) that passed the most stringent PCAWG quality-control criteria: 1 or more somatic structural variants were detected in 2,429 tumours<sup>8</sup>. As described in an accompanying Article<sup>8</sup>, structural variants were identified using aberrantly mapping and/or split reads in paired-end sequencing data<sup>25</sup>. We used four somatic structural-variant callers<sup>20,25–27</sup>, and the final structural-variant dataset comprised events that were returned by  $\geq 2$  callers, merged by a graph-based consensus method<sup>8</sup>. We consider only somatically acquired structural variants in this analysis, and exclude somatic retrotransposition events. Validation of structural-variant calls was undertaken using both manual inspection and pull-down with resequencing of breakpoints. With these approaches, we estimate the sensitivity of the consensus structural-variant call set to be 90% for true calls generated by any 1 of the 4 callers; specificity was estimated as 97.5%<sup>8</sup>. A mean of 3.22 algorithms of the 4 that we used called each structural variant in the consensus set genome-wide, and this differed little across repetitive elements: the mean for short interspersed nuclear elements was 3.22, and the mean for long interspersed nuclear elements was 3.21.

Because the structural variants from a given cancer are often highly clustered, we grouped rearrangements into clusters on the basis of the



**Fig. 2 | Frequency of structural-variant classes across tumour types.** **a**, Violin plots of density of classified structural-variant categories across patients within each histology group. Tumour type panels are sorted in descending order of the average number of structural-variant breakpoints per sample. Within each tumour type, the frequency distribution (y axis) of different structural-variant categories (x axis) across patients is shown as a density: regions of highest density have the greatest width of shaded area. In each panel, the number of patients is indicated at the top right. AdenoCA, adenocarcinoma; BNHL, B-cell non-Hodgkin lymphoma; ChRCC, chromophobe renal cell carcinoma; CLL, chronic lymphocytic leukaemia; CNS, central nervous system; GBM, glioblastoma; HCC, hepatocellular carcinoma; leiomyo, leiomyosarcoma; medullo, medulloblastoma; MPN, myeloproliferative neoplasm; eso, oesophageal; oligo, oligodendrocytic; panc, pancreatic; piloastro, pilocytic astrocytoma; prost, prostate; RCC, renal cell carcinoma; sarc, sarcoma; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; thy, thyroid. **b**, Per-sample counts of complex (bottom) and classified (top) structural-variant breakpoint junctions for oesophageal adenocarcinoma. **c**, Per-sample counts of complex (bottom) and classified (top) structural-variant breakpoint junctions for ovarian adenocarcinoma.

proximity of breakpoints, the overall number of events in that genome and the size distribution of these events (Supplementary Methods). Essentially, a particular cluster contains structural variants that are significantly closer together than expected by chance, given the overall number and orientation of structural variants in that patient. Alongside the clustering, we computed an in silico library of all possible genomic configurations that result from sequential simple structural variants (deletions, tandem duplications, inversions, translocations, and chromosome duplications or losses), to a depth of five rearrangements. We could then compare the genomic configuration of each observed cluster of structural variants against the library to determine how it might have arisen.

This methodology has the advantage that breakpoint junctions are classified according to the wider genomic context in which they occur. This means that, for example, true deletions will be identifiably different from breakpoint junctions that happen to have a deletion-type

orientation but arise within (for instance) a chromothripsis event of markedly different mechanism and properties. Over half the breakpoint junctions that we observed arise within clusters of several or many structural variants (Fig. 2a): removing these junctions from the catalogues of true deletions, tandem duplications and inversions enables a more-precise description of the properties of simple structural variants.

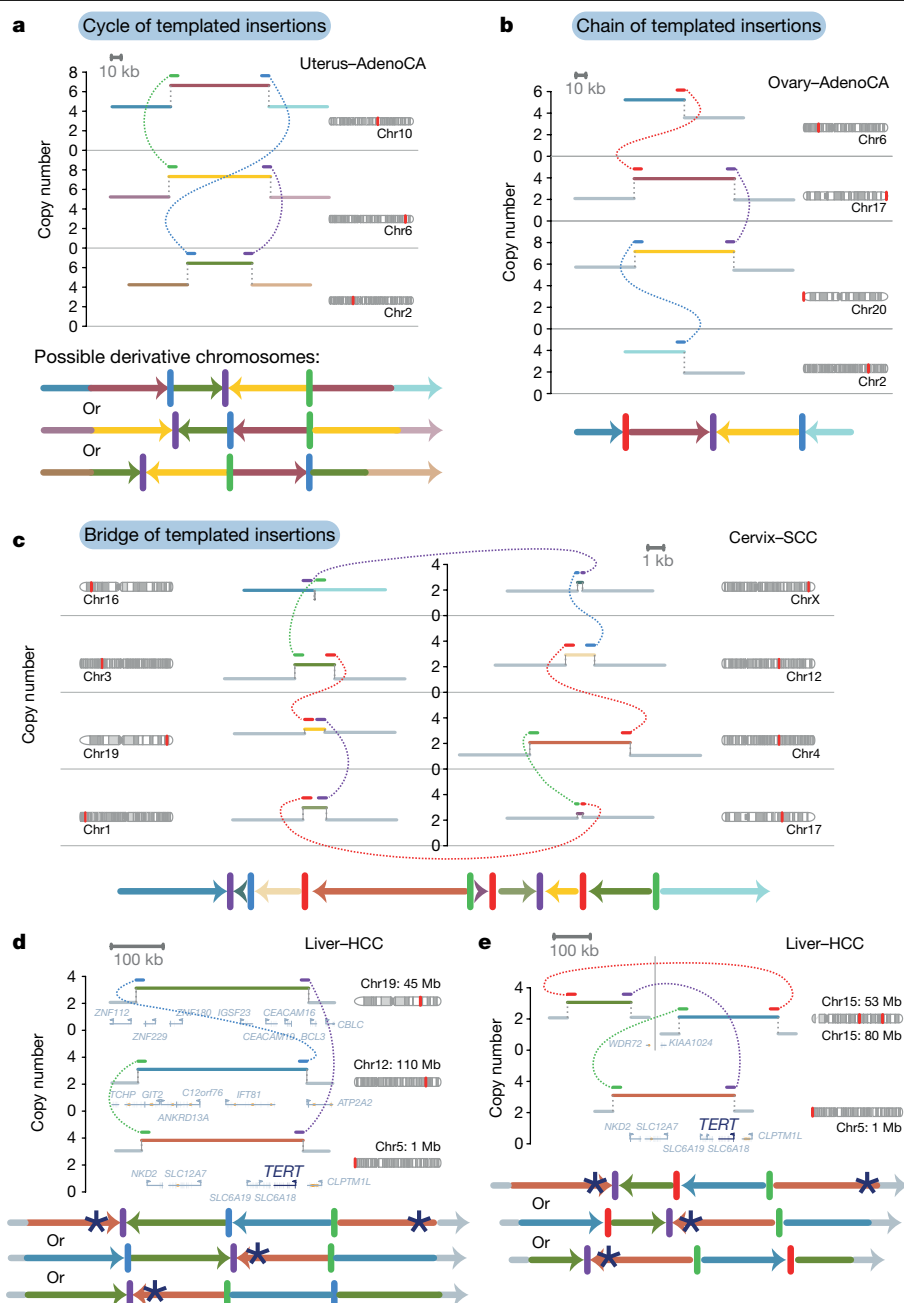
Among the classes of simple structural variants, deletion was the most common, followed by tandem duplication and then unbalanced translocation. Reciprocal translocations and reciprocal inversions were uncommon events (Fig. 2a). There was considerable variability in the overall numbers and distribution of classes of structural variant across tumour types and across patients within a given tumour type (Extended Data Fig. 1). For example, oesophageal adenocarcinomas were characterized by many deletions and a large number of complex clustered rearrangements (Fig. 2b), and ovarian cancers often carried high numbers of tandem duplications and/or deletions with moderate numbers of unbalanced translocations (Fig. 2c).

### Cycles of templated insertions

We next examined clusters that contain 2–10 structural variants. One newly identified configuration consisted of several segments of copy-number gains, typically on different reference chromosomes, linked together through structural variants (Fig. 3, Extended Data Fig. 2). A sequential path through consecutive segments can be formed by following the breakpoint junctions, which suggests that each cluster represents a string of duplicated templates inserted into a single derivative chromosome, probably acquired concurrently. Although it is theoretically possible that the structural variants in such clusters are not phased on the same derivative chromosome or do not occur concurrently, we think this is unlikely for several reasons. First, we found examples of RNA transcripts that spliced together exons separated by two junctions in the structural-variant cluster (Supplementary Fig. 2), which suggests that they are phased on the same derivative chromosome. Second, long-read sequencing data (reported in an accompanying Article<sup>8</sup>) supported the phasing of structural variants that link templated insertions. Third, we found that the clonal fraction of tumour cells tended to be more similar for structural variants within these clusters than for randomly chosen structural variants in each patient (Supplementary Fig. 3), which suggests that they co-occur in evolutionary time. Fourth, the level of copy-number gain for individual segments in the cluster tended to be identical (Fig. 3, Extended Data Fig. 2).

We define three basic categories on the basis of whether or not the string of inserted segments returns to the original chromosome: we term strings of inserted segments that do not return ‘chains’ of templated insertions and those strings that do return ‘bridges’ (which leave a gap on the host chromosome) or ‘cycles’ (which rereplicate a segment on the host chromosome). In the PCAWG dataset overall, we observed 1,467 cycles and 1,275 bridges of templated insertions (Fig. 3a, b, Extended Data Fig. 2). In chains of templated insertions, the string of genomic segments does not return to the chromosome of departure (Fig. 3c, Extended Data Fig. 2) but it is similarly associated with copy-number gains at each templated segment. There were 285 instances of such chains in the dataset, commonly manifesting as unbalanced translocations joined through one or more intermediary templated insertions.

Most templated insertion events involve only two breakpoint junctions, but this can extend to three, four or more linked rearrangements (Extended Data Fig. 3a). The longest such event—from a cervical squamous cell cancer—had seven templated insertions strung together on an eighth host chromosome (Fig. 3c; other examples of long templated insertion events are shown in Extended Data Fig. 3).



**Fig. 3 | Chains, cycles and bridges of templated insertions. a–c**, Examples of a typical cycle (a), chain (b) and bridge (c) of templated insertions. The estimated copy-number profile is shown as in Fig. 1, with structural variants shown as dotted arcs linking two copy-number segments. The derivative

### Templated insertions that affect *TERT*

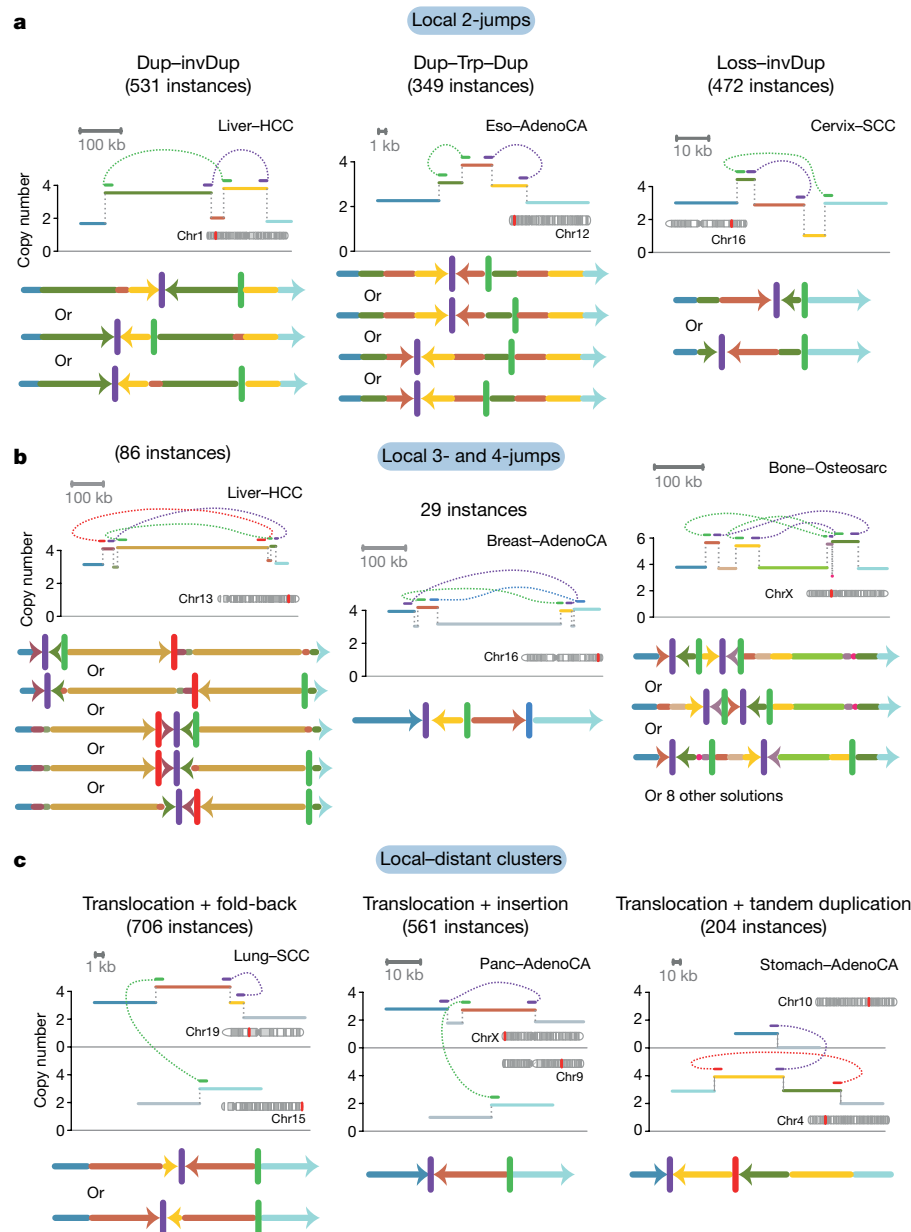
Structural variants drive tumour development through their effects on cancer genes, whether by altering gene copy number, disrupting tumour-suppressor genes, creating fusion genes or juxtaposing the coding sequence of one gene with the regulatory apparatus of another. We found that many liver cancers had cycles of templated insertions that affect *TERT* (Fig. 3d, e, Extended Data Fig. 4). Point mutations in the *TERT* promoter are present in 54% of liver cancers, and a further 5–10% of liver cancers have structural variants that activate the gene<sup>28</sup>. Of the 30 patients with liver cancer that had structural variants that affect *TERT*, we find that 10 of these variants were templated insertion events (mostly cycles). All of these events duplicated the entire

chromosome(s) that could explain the copy-number and structural-variant profile is shown below. **d, e**, Cycles of templated insertions that affect the *TERT* gene, in two hepatocellular carcinomas. *KIAA1024* is also known as *MINARI*.

*TERT* gene and linked it to duplications of whole genes, fragments of genes or regulatory elements from elsewhere in the genome, and led to increased expression of *TERT* (Extended Data Fig. 4e). Thus, this particular rearrangement process is distinctive for the precision with which cancer copy-and-pastes normally disparate functional elements of its genome together without wholesale instability.

Tumour-suppressor genes were also inactivated by templated insertions (Extended Data Fig. 5). For example, among many straightforward deletions, *RB1* was hit by cycles of templated insertions, a templated insertion with deletion and one instance of the linked, inverted duplications detailed in ‘Local *n*-jumps and local-distant clusters’. These events typically generated duplications of internal exons in *RB1* and/or





**Fig. 4 | Examples of clusters of 2–5 rearrangements seen in human cancers. a,** Structures created by two local rearrangements that cannot easily be explained by simple structural-variant classes (which we call local 2-jumps). The estimated copy-number profile is shown as in Fig. 1, with structural variants shown as dotted arcs linking two copy-number segments. Possible configurations of the derivative chromosome are shown below; multiple

solutions are possible for each example. Dup, duplication; invDup, duplication linked by inverted rearrangement; trp, triplication. **b,** Structures created by 3–4 local rearrangements that cannot easily be explained by simple structural-variant categories. **c,** Structures created by one local rearrangement and one rearrangement that reaches elsewhere in the genome (local-distant clusters).

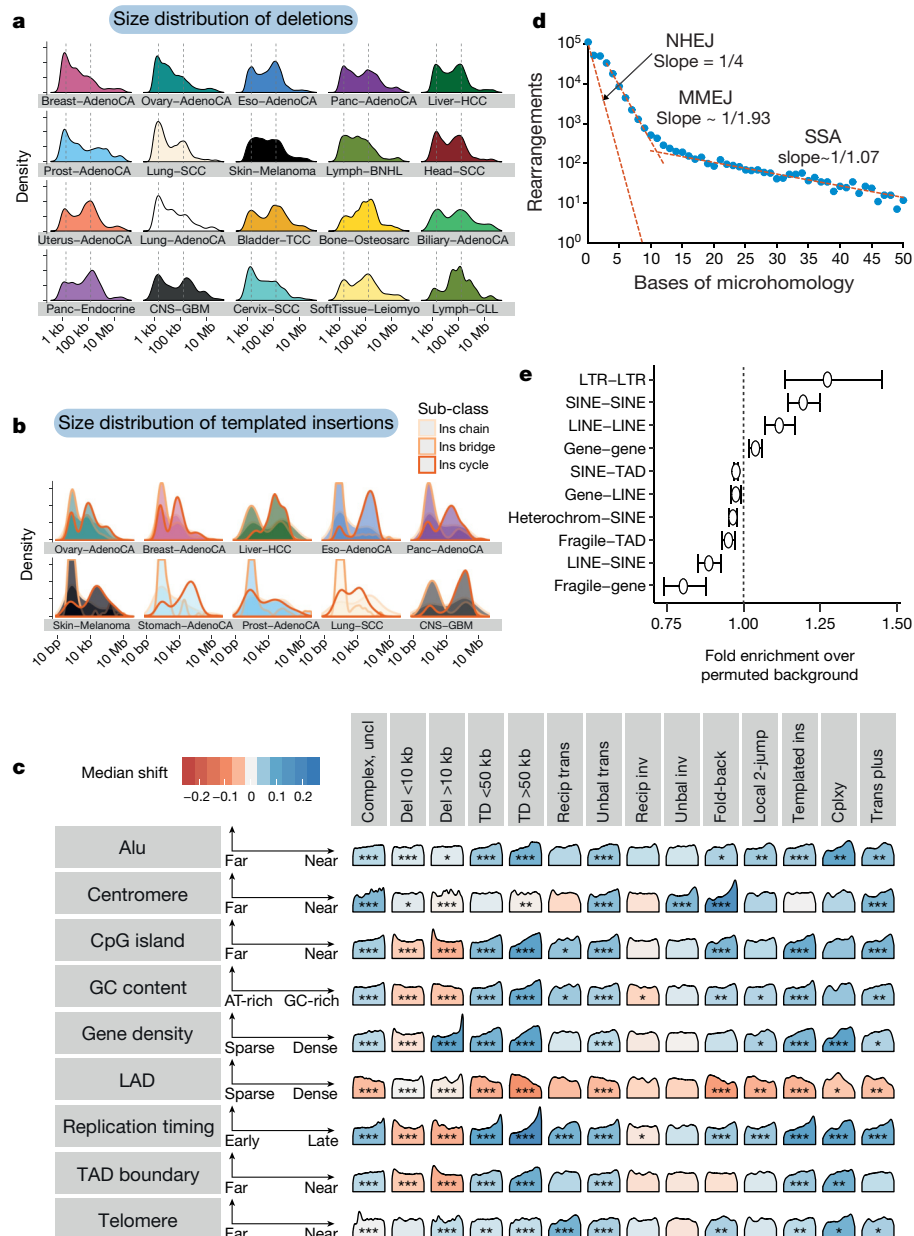
insertions of exons from other genes, all of which presumably rendered a non-functional transcript.

### Local *n*-jumps and local-distant clusters

Many clusters of 2–10 structural variants in the dataset were confined to a single genomic region. Of those clusters that comprised two local rearrangements, some had straightforward explanations, such as nested or adjacent tandem duplications. However, many did not have a trivial explanation (Fig. 4a). These included a duplication–inverted-triplication–duplication structure that has previously been observed in germline structural variants<sup>24</sup> (349 instances); a structure of two duplications linked by inverted rearrangements (531 instances); and structures of copy-number loss plus nearby duplication linked by inverted

rearrangements (472 instances). All of these patterns had solutions in which breakpoints were phased to a single derivative chromosome (Fig. 4a), although non-phased solutions are theoretically possible (if unlikely). Beyond clusters of two rearrangements (two-jumps), we also found examples involving three, four or more rearrangements confined to one genomic locale (Fig. 4b). All of these configurations of clusters of structural variants can be phased to a single derivative chromosome, with tightly grouped breakpoints.

Beyond clusters confined to a single genomic region, we found clusters of 2–10 structural variants that combined local jumps with rearrangements that reach into one or more distant regions of the genome (Fig. 4c). Simple examples of these events include unbalanced translocations or large deletions with a locally derived fragment inserted at the breakpoint, but there was also an extensive



**Fig. 5 | Size distribution and genomic properties of classified structural variants.** **a**, Size distribution of deletions per histology group, with tumour types ordered according to total number of events seen. Vertical dashed lines represent the two prominent modes. **b**, Size distribution of segments of templated insertion per histology group. For each tumour type, the three distributions for cycles, bridges and chains of templated insertions are superimposed. Ins, insertion; Del, deletion; Inv, inversion; TAD, topologically associated domain; TD, tandem duplication; Trans, translocation; Unbal, unbalanced; Recip, reciprocal; LAD, lamina-associated domain; cplx, complex; cplx, complex clusters unclassified; cplx, chromoplexy; del, deletion; inv, inversion; ins, insertion; LAD, lamina-associated domain; recip, reciprocal; TAD, topologically associated domain; TD, tandem duplication; trans, translocation; unbal, unbalanced. **d**, Rearrangement counts as a function of bases of junction microhomology, fit to three linear functions consistent with different formation mechanisms. NHEJ, non-homologous end joining; MMEJ, microhomology-mediated end joining; SSA, single-strand annealing. **e**, Enrichment or depletion of breakpoint junctions between regions of the genome with particular annotations, compared with a permuted background that preserves breakpoint positions but swaps breakpoint partners. Centre points are the mean fold change over the permuted background; error bars represent three s.d. Analysis is based on a sample size of 2,559 genomes containing structural variants. LTR, long terminal repeat; SINE, short interspersed nuclear element; LINE, long interspersed nuclear element; heterochrom, heterochromatin.

range of more-complex patterns. In some cases, the source of the inserted fragment was distal to the major break, and the structural variant could feasibly result from several concurrent DNA breaks in close spatial proximity to the capture of a short DNA fragment during repair (cut-and-paste). In other cases, the origin of the inserted fragment

was proximal to the major break and associated with a gain in copy number. This pattern is difficult to explain by a cut-and-paste mechanism, because the copy-number gain implies the inserted segment was a duplicate of the original template rather than a separated fragment redistributed from its original locus. Instead, a copy-and-paste

mechanism may be the more parsimonious explanation for these events.

A comparison of local footprints linked together through distant rearrangements revealed a strong connectivity of footprints with the same or similar structure, often enriched tenfold or more than expected by chance (see 'Footprint connectivity analysis' in Supplementary Results). The reasons for this are unclear, but it may reflect innate structural symmetry introduced through the generation or the resolution of rearrangements, or through the repeated action of a mechanism that imparts consistent structural motifs.

## Copy-and-paste patterns of clusters

The diverse patterns of 2–10 clustered structural variants (Figs. 3, 4) share important morphological features: (1) genomic configurations that can be phased to a single derivative chromosome; (2) low-level gains in copy number, especially duplications and triplications; (3) a high frequency of inverted rearrangements in addition to noninverted rearrangements; (4) occurrence on a chromosome background with similar average copy number to the tumour overall; and (5) tight proximity of breakpoints within the local footprint (typically <1 Mb).

Using our *in silico* library of genomic configurations, we could define all possible routes by which sequential structural variants could generate these structures through the classically defined repertoire of deletion, tandem duplication, inversion and translocation (Supplementary Fig. 4). These routes typically would require implausible machinations of chromosomes (Supplementary Results). In particular, the high prevalence of inverted breakpoint junctions and local copy-number gains is difficult to recreate using sequential simple rearrangements. Simple inversion events are uncommon in cancers (Fig. 1d) and they tend not to generate copy-number gains, except through breakage–fusion–bridge cycles: these latter also cause terminal deletions<sup>2</sup>, which are not seen in the events discussed here.

If these events cannot be satisfactorily explained by sequential simple rearrangements, another possible explanation is a complex cut-and-paste mechanism such as chromothripsis, chromoplexy or repeated breakage–fusion–bridge cycles. However, the patterns of the 2–10 clustered structural variants do not fit with these processes either (Supplementary Results). Although chromothripsis with copy-number gain has previously been described<sup>3,11,19,22</sup>, the resulting copy number and rearrangement patterns have different properties to those we observed. Chromoplexy, in which chromosome breaks lead to a balanced interchange at multiple breakpoint junctions<sup>5,20</sup>, typically generates unphased solutions. Repeated breakage–fusion–bridge cycles tend to cause high-level copy-number gains associated with inverted, fold-back rearrangements<sup>1,2</sup>, unlike the structures reported here.

Instead, we believe that many of these locally complex clusters of structural variants with low-level copy-number gains are generated in a single event by a copy-and-paste process. That is, the copying of genomic templates is an intrinsic aspect of the structural variation process in these events, with the extra copies being inserted in the resulting derivative chromosome. If the genomic templates all originate locally, we would observe local *n*-jumps (such as in Fig. 3a, b) with a tight clustering of breakpoints, phased solutions, frequent copy-number gains and a mix of inverted and noninverted breakpoint junctions. If the original templates for the copied segments derive from across the genome, chains, cycles and bridges of templated insertions would arise (Fig. 2).

## Genomic properties of structural variants

The size of tandem duplications and deletions followed complex—often multimodal—distributions across tumour types (Fig. 5a, Extended Data Fig. 6a). However, as previously reported<sup>6,29</sup>, individual patients tend to have a simpler—usually unimodal—distribution of deletions or tandem duplications (Extended Data Fig. 6b), which implies that the

complexity seen in a given tumour type results from combining samples with different profiles. The sizes of individual fragments in templated insertion events were also distinctly multimodal, with varying peak heights across tumour types (Fig. 5b). When correlating template sizes within a given event, two patterns emerged: one in which template sizes were closely correlated with one another, and one in which a small (<1 kb) template was linked with one of any size (Extended Data Fig. 7a, b). Likewise, the sizes of segments within a given local two-jump event showed moderately strong correlations with one another (Extended Data Fig. 7c).

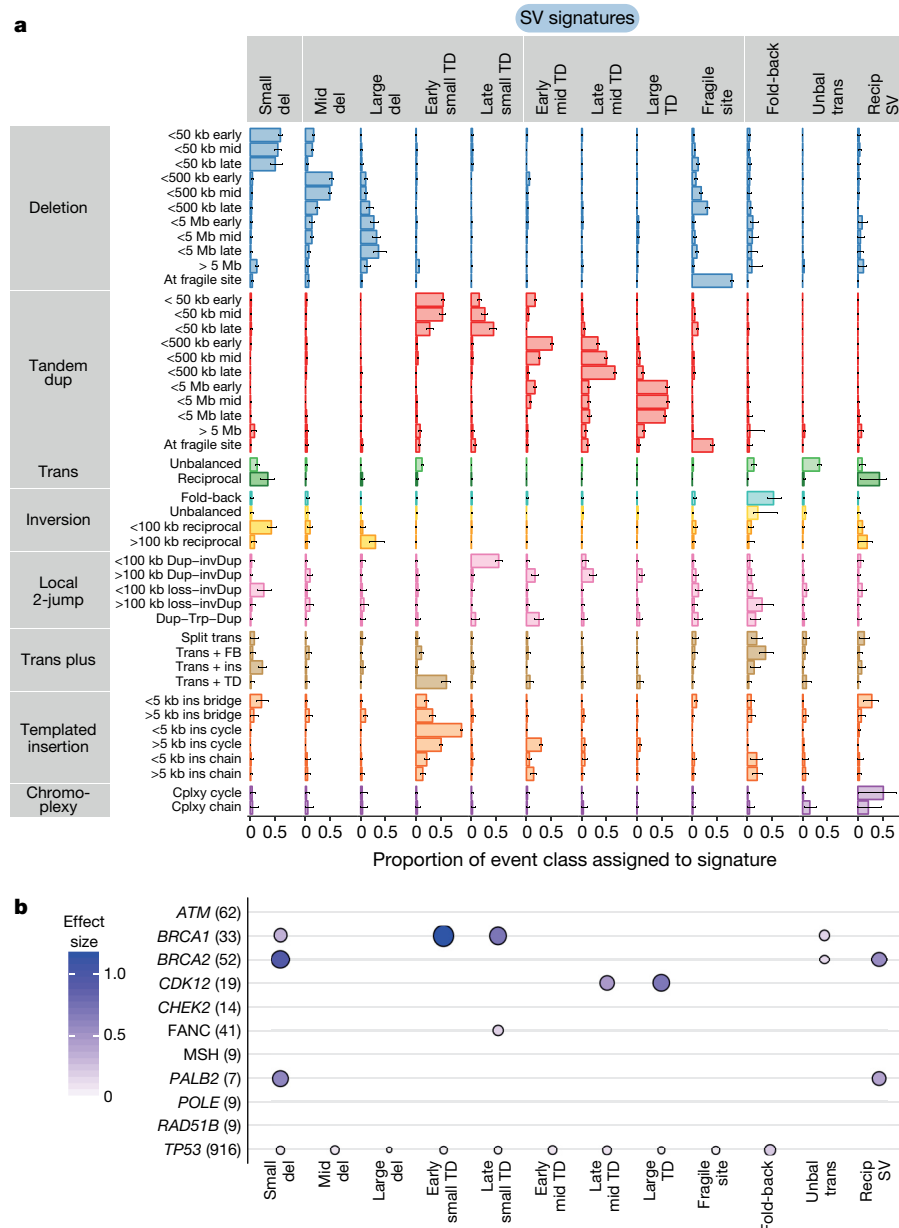
A number of genomic properties (such as replication timing, transcriptional activity and chromatin state) influence the density of point mutations<sup>30,31</sup> and copy-number alterations<sup>32</sup>, but how this relates to individual classes of structural variant is unclear. From the literature, we compiled a library of the genome-wide distribution of 38 features including replication timing, GC content, repeat density, gene density and distance to G-quadruplex motifs, among others. Replication timing had the strongest association with the occurrence of structural variants; deletions are enriched in late-replicating regions, and tandem duplications and unbalanced translocations occur preferentially in early-replicating regions (Fig. 5c, Extended Data Fig. 8). For individual patients with high numbers of deletions or tandem duplications, we observed notable heterogeneity in the distribution of these structural variants according to replication timing: some had events that occurred predominantly in late-replicating regions, others had events that occurred exclusively in early-replicating regions, and in others events were distributed more evenly (Supplementary Fig. 5). Regions of active chromatin and increased gene density correlated positively with the rate of rearrangement.

A structural variant requires DNA repair pathways to join two sequences together, and several repair mechanisms are available to somatic cells. Some require sequence homology between the two ends, and others can operate to join non-homologous sequences. As previously reported<sup>2,25,33</sup>, we find across the PCAWG data that many structural variants do not have sequence homology at the breakpoint junction (Fig. 5d) and therefore arise through non-homologous end joining. Nonetheless, a sizable fraction of structural variants has more microhomology than expected by chance, with an apparently bimodal distribution of microhomology lengths. One set of structural variants has 2–7 bp of microhomology, probably generated by microhomology-mediated end joining, and a second set of structural variants has 10–30 bp of microhomology, probably generated through single-strand annealing or other forms of homologous recombination (including microhomology-mediated break-induced replication). Repetitive sequences in the genome, such as short and long interspersed nuclear elements, are the likely substrate of such structural variants, and we find enrichment for structural variants joining such elements (Fig. 5e, Supplementary Fig. 6).

## Signatures of structural variation

The heterogeneous spectrum of point mutations across cancers can be reconstructed from the differential action of a relatively limited repertoire of mutational processes, each with a characteristic signature<sup>34</sup>. The differences across patients in the size distribution of tandem duplication and deletion—together with the widely varying frequency and patterns of structural variant across tumour types and genome topology—suggested that we could similarly learn such correlations across individual classes of structural variant.

We divided the set of structural variants of each patient into mutually exclusive categories. We split the most frequent classes of simple structural variant (deletions and tandem duplications) into 11 categories according to size, replication timing and occurrence at fragile sites. Other configurations of structural variants and copy-number changes seen more than 50 times in the cohort were included as further



**Fig. 6 | Structural-variant signatures in human cancers. a**, The 12 most distinctive structural-variant signatures extracted by the Bayesian hierarchical Dirichlet process algorithm, run on a sample size of 2,559 genomes containing structural variants. Here the lengths of the bars represent the estimated proportion of each event class assigned to each signature (rows sum to one); the black line segments represent the 95% posterior interval for bar length from the Markov chain. FB, fold-back; mid, mid-sized. **b**, Association of pathogenic mutations (germline and somatic combined) in key DNA repair genes with structural-variant signatures. The sample size of patients who have

pathogenic variants in the specific genes assessed is shown in brackets after each gene label (y axis). Hypothesis tests and effect sizes for each gene are derived from linear models for signature intensity after correction for histology. Significant associations from two-sided tests with correction for multiple hypothesis testing are shown. The colour and size of the points represent the estimated effect sizes. MSH refers to *MSH2*, *MSH3*, *MSH4* and *MSH6*, genes in the mismatch repair pathway; FANCA refers to genes associated with Fanconi anaemia, namely *FANCA*, *FANCC*, *FANCD2*, *FANCE*, *FANCF*, *FANCG*, *FANCI*, *FANCL* and *FANCM*.

categories, including cycles, chains and bridges of templated insertions (also split by size), local *n*-jumps and local-distant clusters.

We applied two methods for signature discovery, which yielded comparable results. We identified 16 structural-variant signatures: the 12 most prevalent of these signatures are shown in Fig. 6a. Signature extraction on the cohort randomly split into two halves identified ten highly correlated signatures (Supplementary Fig. 7), which closely matched the signatures called in the full cohort despite the lower power. Three signatures of deletions emerged, split by size: the signature of small (<50-kb) deletions included small reciprocal inversions and the signature of large (>500-kb) deletions included large reciprocal inversions. This implies that the frequencies of deletions and reciprocal

inversions are correlated across the cohort, and both follow similar size distributions within an individual patient.

We identified five signatures of tandem duplications, split by size and replication timing. Cycles, bridges and chains of templated insertions were particularly prominent in signatures of early-replicating tandem duplications, whereas local two-jump structures were more closely associated with late-replicating tandem duplications. All of these patterns exemplify the copy-and-paste concept, in which extra copies of genomic templates are produced and inserted as an integral feature of the structural-variant process.

Another signature was characterized by deletions and tandem duplications at chromosomal fragile sites<sup>35</sup>. Tandem duplications were more

prominent at the edges of the fragile site, and deletions were concentrated in the centre (Extended Data Fig. 9a, b). The size range of fragile site deletions peaked at around 100 kb, similar to the larger deletion signature, whereas the rarer fragile-site tandem duplications showed no strong size peak (Extended Data Fig. 9c). Sites of fragility varied extensively across tumour types (Extended Data Fig. 9d).

Unbalanced translocations comprised their own signature, which suggests that they derive from a distinct rearrangement process in cancer genomes. A further signature comprised both the fold-back inversions that are a hallmark of breakage–fusion–bridge cycles and similar structures such as translocations adjacent to fold-back inversions. Finally, there was a signature of balanced rearrangements, including reciprocal translocations and chromoplexy clusters<sup>5</sup>. This signature probably arises from several double-stranded DNA breaks (potentially occurring in interphase), in which both sides of the break are incorrectly repaired through ligation to other, simultaneously broken regions of the genome.

### DNA repair genes and tumour type

We grouped annotations of pathogenic germline variants and somatic driver mutations in DNA-repair genes across the cohort<sup>8</sup>, correlating their presence with activity of the structural-variant signatures (Fig. 6b). As previously described for breast and ovarian cancers<sup>6,29</sup>, *BRCA1* mutations are significantly associated with small tandem duplication signatures, the mechanistic basis of which is increasingly well understood<sup>10</sup>. As previously described<sup>6,36</sup>, *CDK12* variants predicted signatures of mid-sized-to-large tandem duplications. *BRCA2* variants correlated with small deletions, as expected from previous work<sup>29</sup>, and also with the reciprocal structural-variant signature that includes chromoplexy. *PALB2* variants showed the same correlations with signatures of small deletions and reciprocal structural variants as does *BRCA2*: *PALB2* colocalizes with, stabilizes and assists *BRCA2* during homologous recombination<sup>37</sup>, so we might have predicted that inactivation of either gene would lead to a similar structural-variant signature. These associations between driver mutations and structural-variant signatures were consistently evident across many types of tumour (Extended Data Fig. 10).

The structural-variant signatures showed considerable heterogeneity in their activity across tumour types and among patients within a given tumour type (Supplementary Fig. 8). Tumours of the gastrointestinal tract—including colorectal and oesophageal adenocarcinomas—showed high rates of the fragile-site signature. Prostate cancer was notable for the prevalence of the chromoplexy signature, as previously reported<sup>5,20</sup>, and squamous cell carcinomas of the lung were characterized by the fold-back inversion signature.

We assessed how classes of structural variant altered known cancer genes (Supplementary Table 1). Some cancer genes acquire oncogenic potential only with specific structural events, such as fusion genes or enhancer hijacking. Not surprisingly, these genes typically showed little variability in which classes of structural variant could generate such events (Extended Data Fig. 11a–c)—although there were exceptions. The *TMPRSS2-ERG* fusion gene of prostate cancer, for example, was generated by a range of processes (including simple deletions, chromoplexy and chromothripsis), all of which are prevalent signatures in this tumour type (Extended Data Fig. 11d–f).

Tumour-suppressor genes and recurrently amplified genes showed more variability in which types of structural variant were observed, and these were shaped by signatures active in the relevant tumour types. For example, the tumour-suppressor genes, *PTEN* and *RADS1B*, which are commonly inactivated in breast and ovarian cancers, were often targeted by tandem duplications generating out-of-frame exon duplications (Extended Data Fig. 12a, b). By contrast, deletions were the predominant events that inactivated *SMAD4* and *CDKN2A*, in keeping with their prevalence in cancers of the gastrointestinal tract

(Extended Data Fig. 12c, d). *MYC*, one of the most commonly amplified genes across all types of cancer, showed considerable diversity in the mechanisms of its rearrangement: nested tandem duplications in breast cancer, translocations or chromoplexy with *IGH* in lymphoma, as well as chromothripsis, cycles of templated insertions, local *n*-jumps and local–distant clusters in other types of tumour (Extended Data Fig. 13).

### Discussion

We have described the patterns and signatures of structural variation in a large cohort of uniformly analysed cancer genomes. A major grouping of patterns in structural variants that emerges from our study is one in which extra copies of genomic templates are inserted during the rearrangement process. This includes simple events such as tandem duplications, as well as a range of more-complex events with duplications and triplications that are rearranged locally as well as inserted distantly. Our signature analysis grouped a large proportion of these more-complex events together with tandem duplications, which suggests that they represent a continuum of processes that share underlying properties. A replication-based mechanism has previously been proposed to explain local two-jumps<sup>4,23,24</sup>, in which stalled replication forks or other DNA lesions cause the DNA polymerase to switch templates and continue replication in a new location. Studies in experimental models are now revealing that a wide range of mechanisms and DNA lesions can result in templated insertions: these mechanisms include tandem duplications in *BRCA1* deficiency<sup>10</sup>, translocations with templated insertions caused by dysregulated strand invasion<sup>38</sup> and distant templated insertions in the absence of replication helicases<sup>39</sup>.

Genomic instability in cancer is not a single phenomenon. Instead, many different mutational processes can act to restructure the genome and, in doing so, generate a notably flexible array of possible structures. Any given tumour draws on a subset of the available processes, shaped by the cell of origin, germline predisposition and other, unknown, factors: selection then does the rest, promoting the clone that has chanced on the structure that increases its potential for self-determination.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1913-9>.

1. Bignell, G. R. et al. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17**, 1296–1303 (2007).
2. Campbell, P. J. et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
3. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
4. Lee, J. A., Carvalho, C. M. & Lupski, J. R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).
5. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
6. Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. *Cancer Cell* **34**, 197–210 (2018).
7. Liu, P. et al. An organismal CNV mutator phenotype restricted to early human development. *Cell* **168**, 830–842 (2017).
8. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
9. Zhang, C.-Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
10. Willis, N. A. et al. Mechanism of tandem duplication formation in *BRCA1*-mutant cells. *Nature* **551**, 590–595 (2017).
11. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and kataegis induced by telomere crisis. *Cell* **163**, 1641–1654 (2015).
12. Ly, P. et al. Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. *Nat. Genet.* **51**, 705–715 (2019).
13. Ghezraoui, H. et al. Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. *Mol. Cell* **55**, 829–842 (2014).



14. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1965-x> (2020).
15. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations in cancer. *Nature* <https://doi.org/10.1038/s41586-020-1970-0> (2020).
16. Akdemir, K. C. et al. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0564-y> (2020).
17. Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0562-0> (2020).
18. Cortes-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0576-7> (2020).
19. Li, Y. et al. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98–102 (2014).
20. Berger, M. F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
21. Crasta, K. et al. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**, 53–58 (2012).
22. Rausch, T. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71 (2012).
23. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
24. Carvalho, C. M. B. et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43**, 1074–1081 (2011).
25. Campbell, P. J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
26. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
27. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
28. Totoki, Y. et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat. Genet.* **46**, 1267–1273 (2014).
29. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
30. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
31. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
32. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.* **29**, 1103–1108 (2011).
33. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
34. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* <https://doi.org/10.1038/s41586-020-1943-3> (2020).
35. Lukusa, T. & Fryns, J. P. Human chromosome fragility. *Biochim. Biophys. Acta* **1779**, 3–16 (2008).
36. Popova, T. et al. Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. *Cancer Res.* **76**, 1882–1891 (2016).
37. Xia, B. et al. Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol. Cell* **22**, 719–729 (2006).
38. Piazza, A., Wright, W. D. & Heyer, W. D. Multi-invasions are recombination byproducts that induce chromosomal rearrangements. *Cell* **170**, 760–773 (2017).
39. Yu, Y. et al. Dna2 nuclease deficiency results in large and complex DNA insertions at chromosomal breaks. *Nature* **564**, 287–290 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

#### PCAWG Structural Variation Working Group

Kadir C. Akdemir<sup>15</sup>, Eva G. Alvarez<sup>16,17,18</sup>, Adrian Baez-Ortega<sup>19</sup>, Rameen Beroukhi<sup>3,4,5</sup>, Paul C. Boutros<sup>20,21,22,23</sup>, David D. L. Bowtell<sup>24,25</sup>, Benedikt Brors<sup>26,27,28</sup>, Kathleen H. Burns<sup>29</sup>, Peter J. Campbell<sup>11,12</sup>, Kin Chan<sup>30</sup>, Ken Chen<sup>15</sup>, Isidro Cortés-Ciriano<sup>31,32,33</sup>, Ana Dueso-Barroso<sup>34</sup>, Andrew J. Dunford<sup>35</sup>, Paul A. Edwards<sup>35,36</sup>, Xavier Estivill<sup>37,38</sup>, Dariush Etemadmoghadam<sup>24</sup>, Lars Feuerbach<sup>27</sup>, J. Lynn Fink<sup>34,39</sup>, Milana Frenkel-Morgenstern<sup>40</sup>, Dale W. Garsed<sup>24</sup>, Mark Gerstein<sup>41,42,43</sup>, Dmitry A. Gordenin<sup>44</sup>, David Haan<sup>45</sup>, James E. Haber<sup>8</sup>, Julian M. Hess<sup>3,46</sup>, Barbara Hutter<sup>26,28,47</sup>, Marcin Imielinski<sup>6,9</sup>, David T. W. Jones<sup>48,49</sup>, Young Seok Ju<sup>150</sup>, Marat D.

Kazanov<sup>51,52,53</sup>, Leszek J. Klimczak<sup>54</sup>, Youngil Koh<sup>55,56</sup>, Jan O. Korbel<sup>7</sup>, Kiran Kumar<sup>3</sup>, Eunjung Alice Lee<sup>57,58</sup>, Jake June-Koo Lee<sup>32,33</sup>, Yilong Li<sup>12</sup>, Andy G. Lynch<sup>35,36,59</sup>, Geoff Macintyre<sup>35,36</sup>, Florian Markowetz<sup>35,36</sup>, Iñigo Martincorena<sup>1</sup>, Alexander Martinez-Fundichely<sup>60,61,62</sup>, Matthew Meyerson<sup>3,4,63</sup>, Satoru Miyano<sup>64</sup>, Hirowaki Nakagawa<sup>65</sup>, Fabio C. P. Navarro<sup>66</sup>, Stephan Ossowski<sup>67,68,69</sup>, Peter J. Park<sup>32,33</sup>, John V. Pearson<sup>70,71</sup>, Montserrat Puiggròs<sup>34</sup>, Karsten Rippe<sup>72</sup>, Nicola D. Roberts<sup>1</sup>, Steven A. Roberts<sup>73</sup>, Bernardo Rodriguez-Martin<sup>16,17,18</sup>, Steven E. Schumacher<sup>3,4,5</sup>, Ralph Scully<sup>74</sup>, Mark Shackleton<sup>24,25</sup>, Nikos Sidiropoulos<sup>11</sup>, Lina Sieverling<sup>27,75</sup>, Chip Stewart<sup>3</sup>, David Torrents<sup>34,76</sup>, Jose M. C. Tubio<sup>16,17,18</sup>, Izar Villasante<sup>34</sup>, Nicola Waddell<sup>70,71</sup>, Jeremiah A. Wala<sup>3,4,5</sup>, Joachim Weischenfeldt<sup>11</sup>, Lixing Yang<sup>77</sup>, Xiaotong Yao<sup>8,78</sup>, Sung-Soo Yoon<sup>56</sup>, Jorge Zamora<sup>116,17,18</sup> & Cheng-Zhong Zhang<sup>3,4,63</sup>

<sup>15</sup>University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>16</sup>Department of Zoology, Genetics and Physical Anthropology, University of Santiago de Compostela, Santiago de Compostela, Spain. <sup>17</sup>Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), University of Santiago de Compostela, Santiago de Compostela, Spain. <sup>18</sup>The Biomedical Research Centre (CINBIO), University of Vigo, Vigo, Spain. <sup>19</sup>Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. <sup>20</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>21</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. <sup>22</sup>Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. <sup>23</sup>University of California Los Angeles, Los Angeles, CA, USA. <sup>24</sup>Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. <sup>25</sup>Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. <sup>26</sup>National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany. <sup>27</sup>Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>28</sup>German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>29</sup>Johns Hopkins School of Medicine, Baltimore, MD, USA. <sup>30</sup>Faculty of Medicine, Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Ontario, Canada. <sup>31</sup>Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. <sup>32</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>33</sup>Ludwig Center, Harvard Medical School, Boston, MA, USA. <sup>34</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain. <sup>35</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>36</sup>University of Cambridge, Cambridge, UK. <sup>37</sup>Sidra Medicine, Doha, Qatar. <sup>38</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain. <sup>39</sup>Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. <sup>40</sup>The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. <sup>41</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA. <sup>42</sup>Department of Computer Science, Yale University, New Haven, CT, USA. <sup>43</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. <sup>44</sup>Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. <sup>45</sup>Biomedical Engineering Department, University of California, Santa Cruz, Santa Cruz, CA, USA. <sup>46</sup>Massachusetts General Hospital Center for Cancer Research, Charlestown, MA, USA. <sup>47</sup>Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>48</sup>Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany. <sup>49</sup>Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>50</sup>Korea Advanced Institute of Science and Technology, Daejeon, South Korea. <sup>51</sup>Skolkovo Institute of Science and Technology, Moscow, Russia. <sup>52</sup>A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia. <sup>53</sup>Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia. <sup>54</sup>Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. <sup>55</sup>Center For Medical Innovation, Seoul National University Hospital, Seoul, South Korea. <sup>56</sup>Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. <sup>57</sup>Division of Genetics and Genomics, Harvard Medical School, Boston, MA, USA. <sup>58</sup>Boston Children's Hospital, Boston, MA, USA. <sup>59</sup>School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. <sup>60</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. <sup>61</sup>Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. <sup>62</sup>Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. <sup>63</sup>Dana-Farber Cancer Institute, Boston, MA, USA. <sup>64</sup>The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>65</sup>RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>66</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. <sup>67</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>68</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>69</sup>Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. <sup>70</sup>Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. <sup>71</sup>Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. <sup>72</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>73</sup>School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA, USA. <sup>74</sup>Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, USA. <sup>75</sup>Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. <sup>76</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. <sup>77</sup>Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL, USA. <sup>78</sup>Tri-institutional PhD Program of Computational Biology and Medicine, Weill Cornell Medicine, New York, NY, USA.

# Article

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

A detailed description of the methods used in this paper and many additional results are described in Supplementary Information. Here, we summarize the key aspects of the analysis.

### Generation of the structural-variant call set

The final set of structural variants used in this Article was generated by the Technical Working Group of the PCAWG Consortium and is described in the main PCAWG paper<sup>8</sup>. In brief, four variant callers were used to identify somatically acquired structural variants from matched tumour and germline whole genome sequencing data: SvABA (Broad pipeline), DELLY (DKFZ pipeline), BRASS (Sanger pipeline) and dRanger (Broad pipeline). These were merged into a final call set using a graph-based algorithm to identify overlapping breakpoint junctions across algorithms. Detailed visual inspection of structural-variant calls suggested that a simple approach of accepting all structural-variant calls made by two or more of the four algorithms gave the best trade-off between sensitivity and specificity.

### Structural-variant clustering and annotation

To identify clusters of structural variants, we developed a method for grouping structural variants into clusters and footprints to allow structural and mechanistic inferences to be made systematically. In parallel, we processed the somatic copy-number data and merged it with structural-variant junctions to enable us produce rearrangement patterns from the generated structural-variant clusters and footprints. We produced normalized representations of structural-variant cluster patterns, which enable us to tabulate the number of different cluster and footprint patterns and analyse their features. Finally, we performed manual and simulation-assisted interpretation of the recurrently observed cluster and footprint patterns. The individual steps of the structural-variant classification pipeline are outlined below and detailed in the subsequent subsections: (1) computing the exact breakpoint coordinates from clipped reads; (2) removing redundant 'segment-bypassing' structural variants; (3) merging rearrangement breakpoints with copy-number data to yield structural-variant breakpoint-demarcated, normalized, absolute copy-number data; (4) clustering individual structural variants into structural-variant clusters and footprints; (5) heuristically refining structural-variant clusters and footprints; (6) filtering artefactual fold-back-type structural variants with insufficient support; (7) determining balanced overlapping breakpoints (this step is to distinguish very short templated insertions from mutually overlapping balanced breakpoints); and (8) computing rearrangement patterns and categories.

### Distribution of structural variants across the genome

We divided the hg19 human reference genome (autosomes and chromosome X) into 3,036,315 pixels of 1 kb, and calculated a suite of metrics per pixel to summarize a variety of genome properties with potential relevance to the distribution of rearrangements, as listed in the Supplementary Information. Properties were matched as closely as possible to the tissue of origin for cancer samples from the PCAWG data. All other genome properties were held fixed across all tissues. To test for associations between structural-variant event classes and the library of genome properties, the genome property metrics were compared between real structural-variant positions (randomly choosing one side of each breakpoint junction to reduce dependence between observations) and one million uniform random positions from the callable genome space. To compare the tissue-specific properties, each random position was assigned a random tissue type, drawing from the observed tissue-type distribution in the structural-variant call set.

For each genome property and each event class, the real observations were pooled amongst the random ones, and then rank-transformed and normalized on a scale from 0 to 1. Under the null hypothesis of no event-versus-property association, the ranks of the real observations would follow a uniform distribution. We tested this in each case with a Kolmogorov–Smirnov test then applied a Benjamini–Yekutieli correction for false-discovery rate across the entire suite of tests and set the threshold for significance reporting at 0.01.

### Structural-variant-signature analysis

We used two algorithms for extracting structural-variant signatures. Both used the same input files, comprising a matrix of counts per patient (across all patients) of structural-variant clusters falling into a number of mutually exclusive categories. These categories included the major classes of structural variants, with the more-common events (deletions, tandem duplications and inversions) split by size and/or replication timing. The two algorithms that were used for extracting the signatures were (1) a hierarchical Dirichlet process and (2) non-negative matrix factorization. Further details on the implementation of these algorithms are available in the Supplementary Information.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA PCAWG Consortium are described in an accompanying Article<sup>8</sup> and are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access information that could potentially identify participants, such as germline alleles and the underlying sequencing data, researchers will need to apply to the TCGA data access committee via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC data access compliance office (<http://icgc.org/daco>) for the ICGC portion of the dataset. In addition, to access somatic single-nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

### Code availability

The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at <https://dockstore.org/search?search=pcawg> under the GNU General Public License v.3.0, which allows for reuse and distribution. These are described in detail in an accompanying Article<sup>8</sup>. The code for grouping structural variants into structural-variant clusters and footprints is available at <https://github.com/cancerit/ClusterSV/> (version 1.0). The code for simulating rearrangements can be found at <https://github.com/cancerit/SimSvGenomes> (version 1.0). The code for sampling from the hierarchical Dirichlet process for identification of mutational signatures is implemented as an R package at <https://github.com/nicolaroberts/hdp> (version 0.1.1).

**Acknowledgements** This work was supported by the Wellcome Trust, Pediatric Low-Grade Astrocytoma Fund and the Fund for Innovation in Cancer Informatics. P.J.C. is a Wellcome Trust Senior Clinical Fellow (WT088340MA). We acknowledge the contributions of the many clinical networks across ICGC and TCGA, which provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for the collation, realignment and harmonized variant-

calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

**Author contributions** Y.L., N.D.R., J.A.W. and O.S. contributed equally to this manuscript, undertaking evaluation and curation of structural-variant calls, merging structural-variant call sets from four separate algorithms into a final dataset. Y.L. performed the clustering and classification of structural variants, and identified patterns of rearrangement, with assistance from N.D.R. and M.I. N.D.R. performed the analysis of structural-variant signatures with assistance from Y.L. N.D.R., J.A.W. and O.S. analysed the distribution of structural variants across the genome, with input from J.E.H., E.K., K.K. and S.E.S. S.W. and J.O.K. contributed to the analysis of how germline variants influenced signatures of structural variants. J.W., R.B. and P.J.C. jointly oversaw the project, assisted with data interpretation and wrote the paper, with input from all authors.

**Competing interests** R.B. owns equity in Ampressa Therapeutics; M.M. is the scientific advisory board chair of—and consultant for— Origimed, and receives research funding from Bayer and Ono Pharma, and patent royalties from LabCorp.; J.W. is a consultant for Nference Inc.; C.-Z.Z. is a cofounder and equity holder of Pillar Biosciences, a for-profit company specializing in the development of targeted sequencing assays.

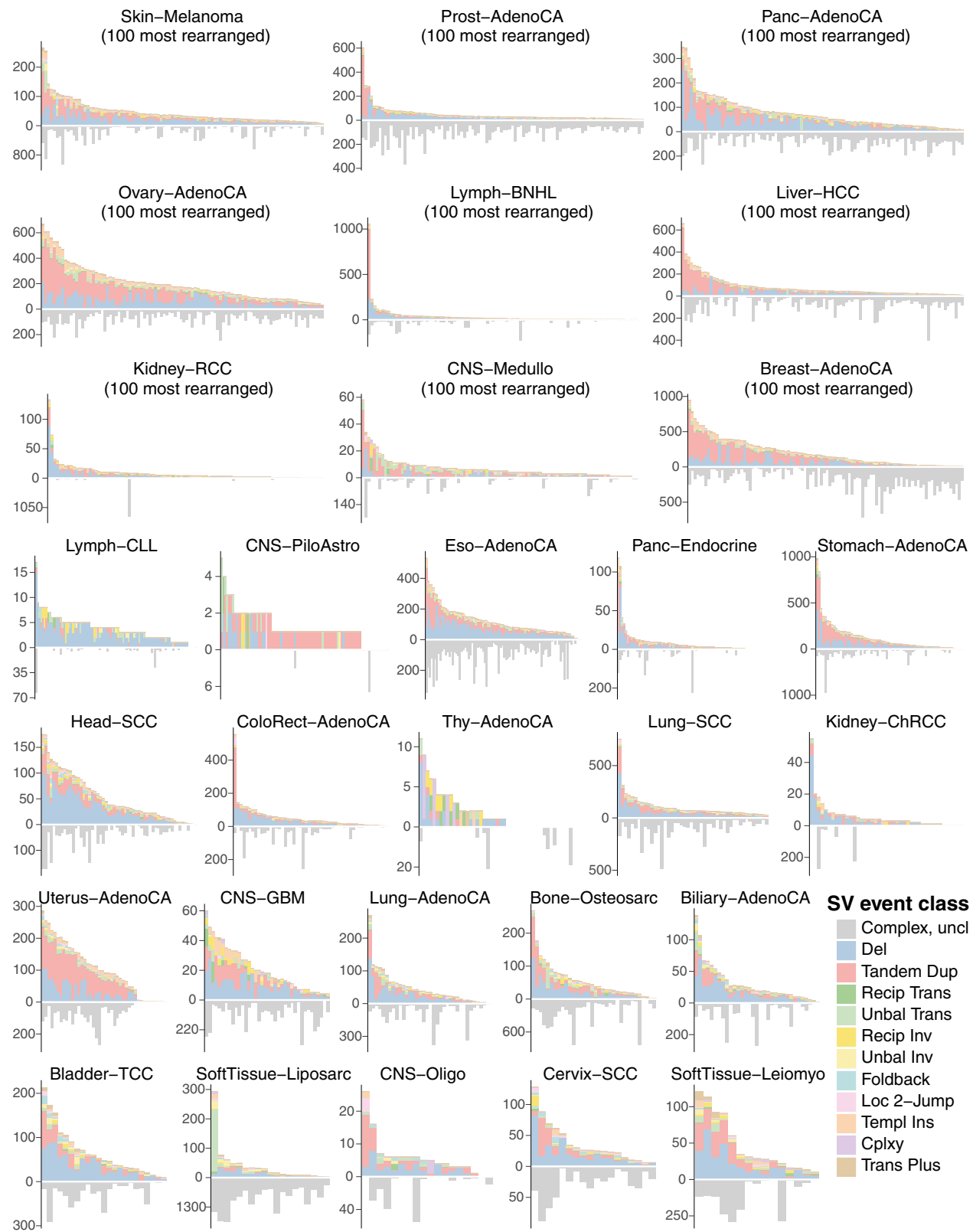
**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1913-9>.

**Correspondence and requests for materials** should be addressed to J.W., R.B. or P.J.C.

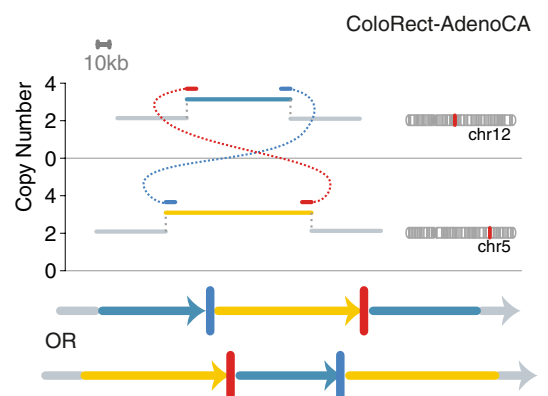
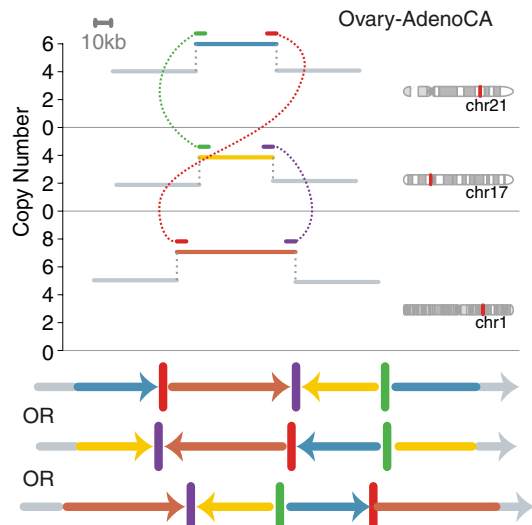
**Peer review information** *Nature* thanks Don Conrad, Ben Lehner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



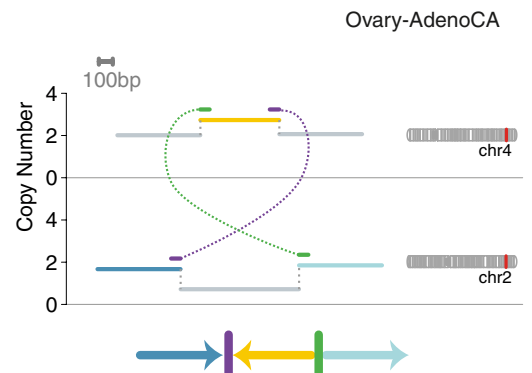
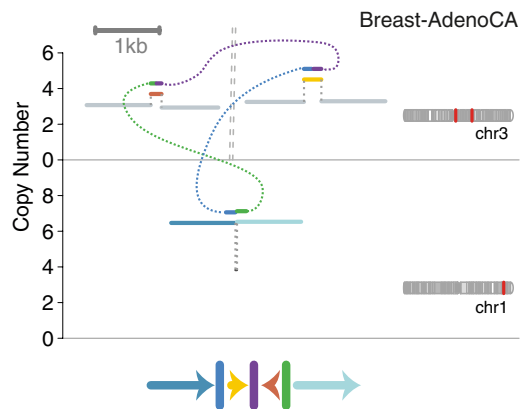
A

**Cycle of templated insertions**



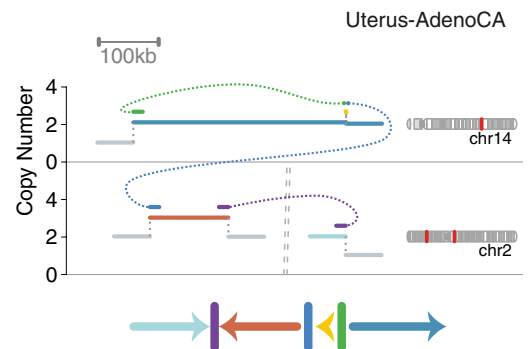
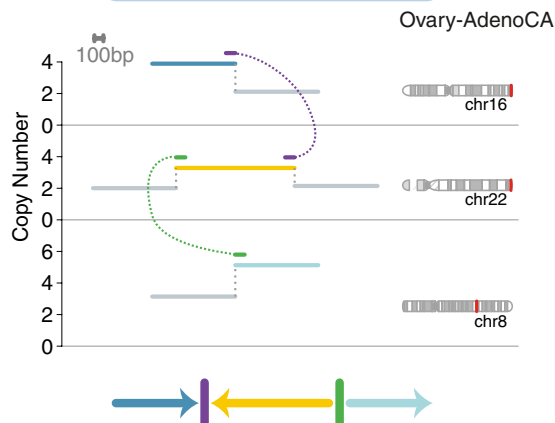
B

**Bridge of templated insertions**



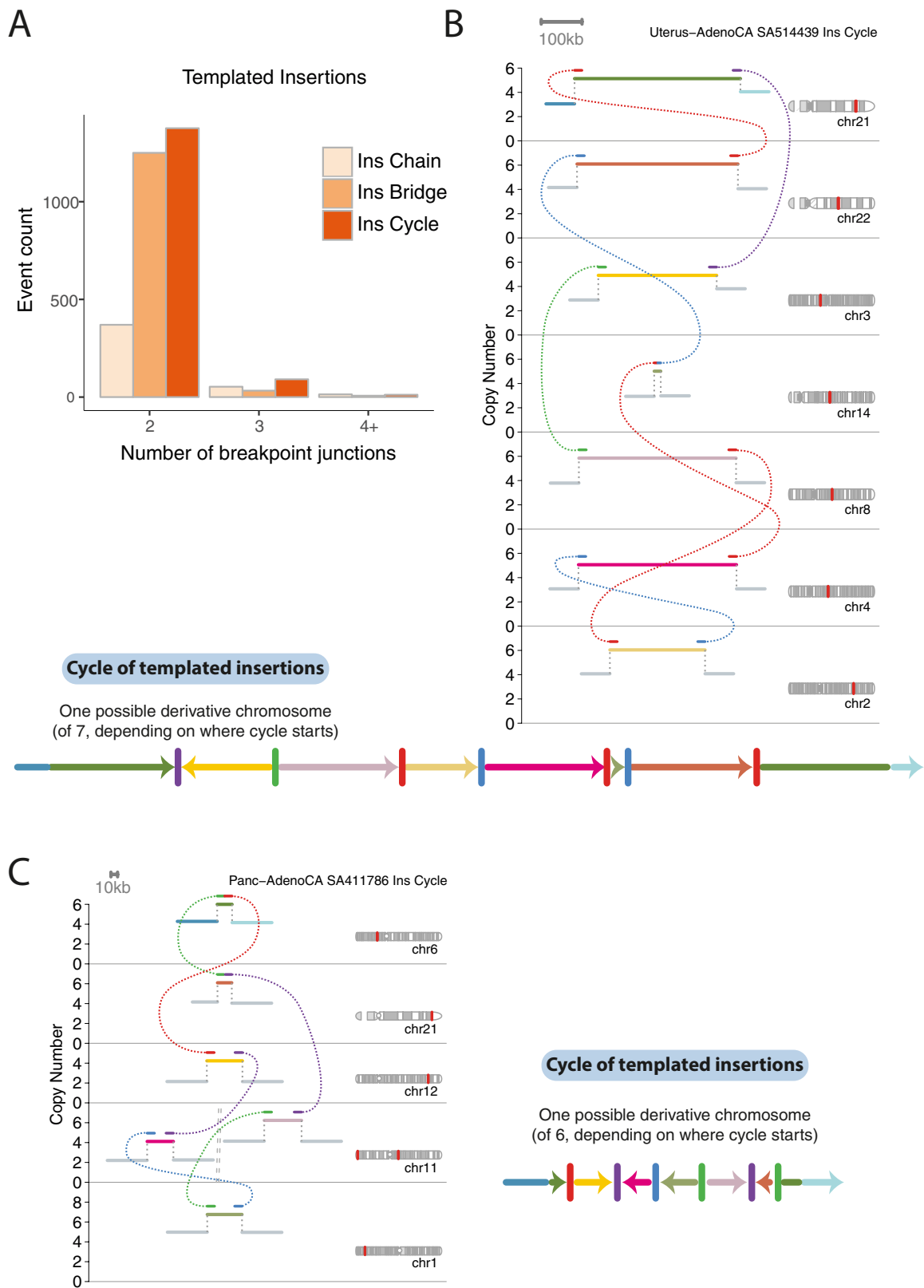
C

**Chain of templated insertions**



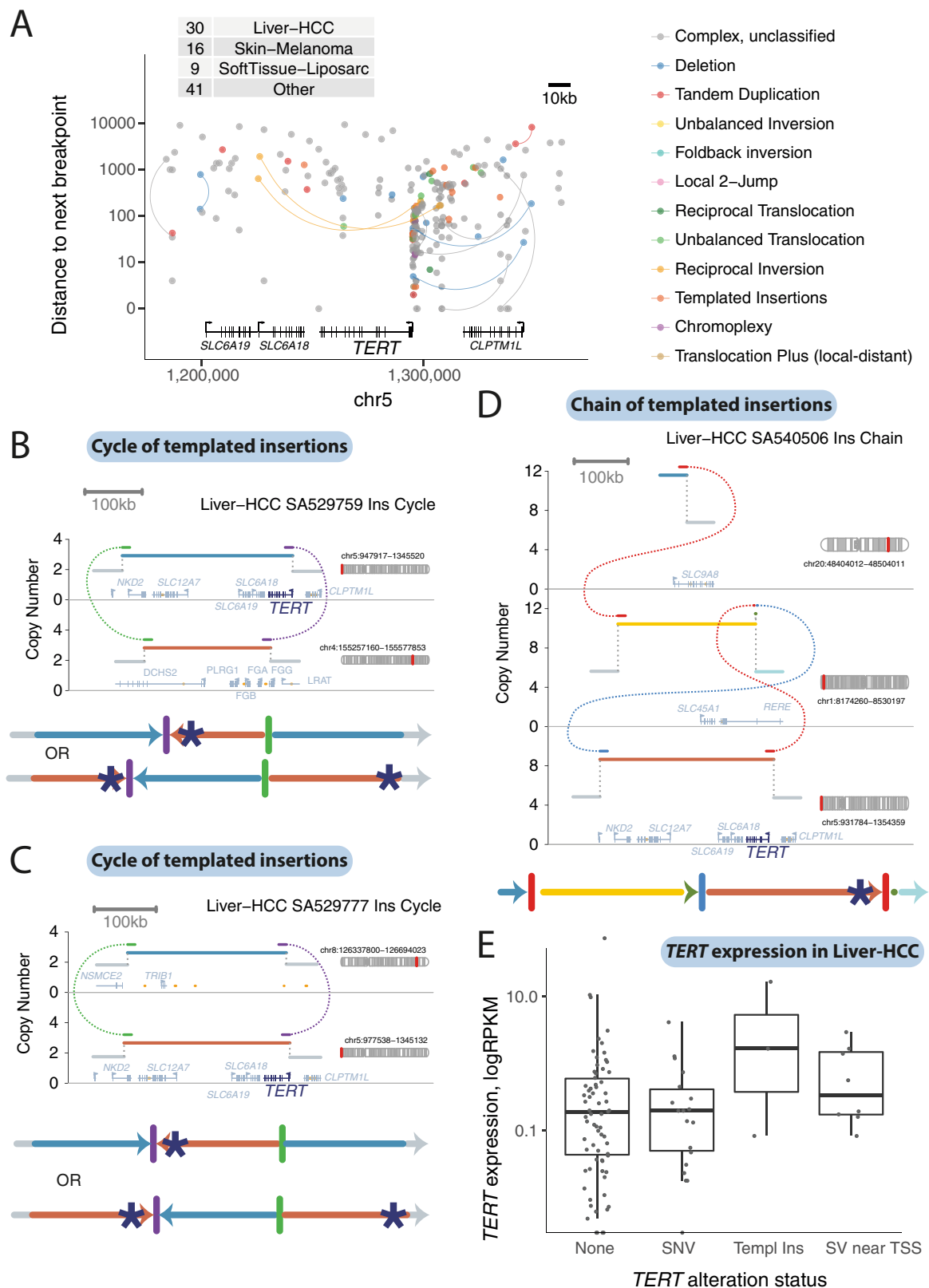
**Extended Data Fig. 2 | Further examples of templated insertion chains, cycles and bridges.** Schematics follow the same structure as in Fig. 3.





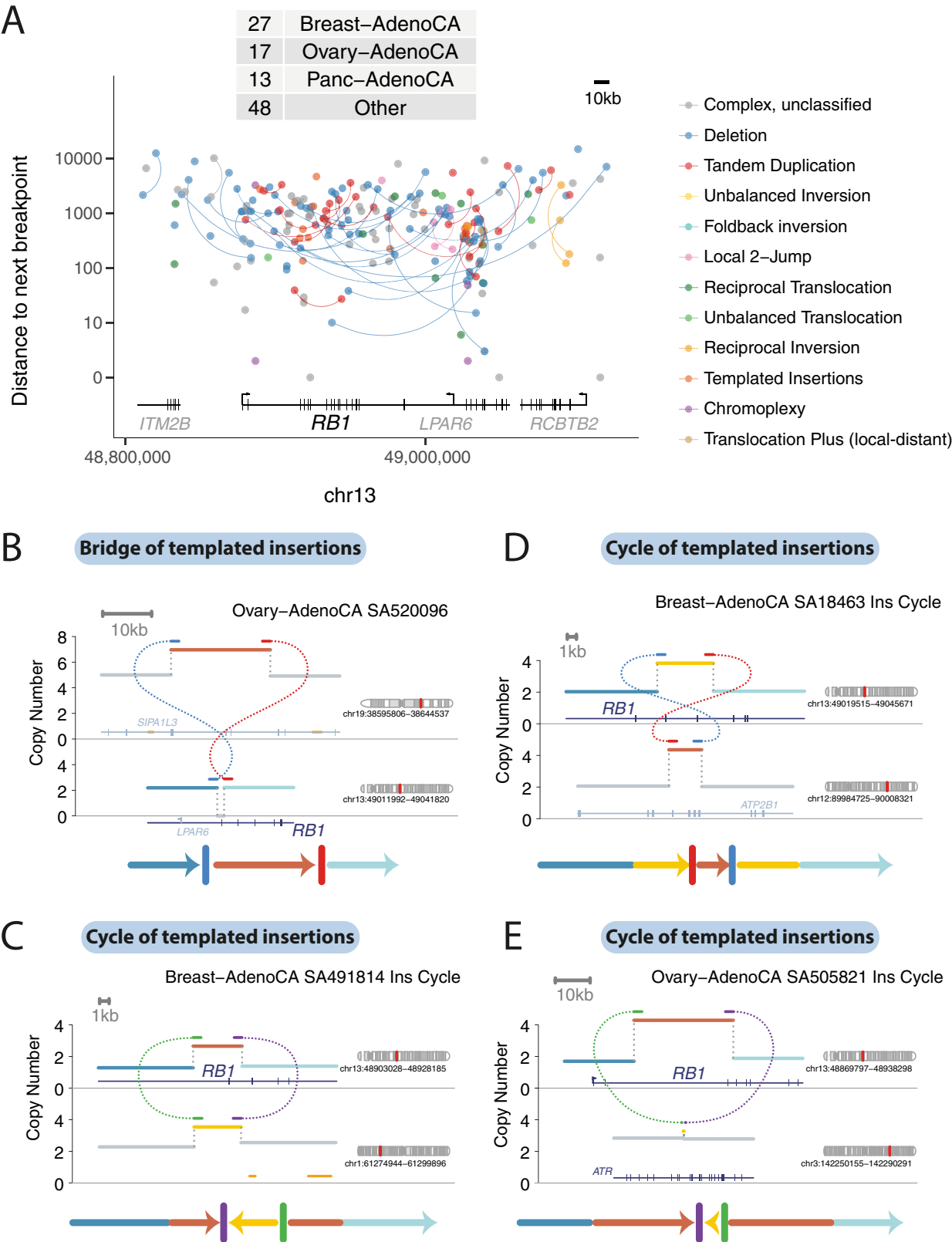
**Extended Data Fig. 3 | Number of breakpoint junctions in cycles, bridges and chains of templated insertions.** **a**, Histogram of numbers of breakpoint junctions in templated insertion cycles, chains and bridges across all samples

in all tumour types in the cohort. **b, c**, Two examples of particularly long cycles of templated insertions in the cohort. Examples are depicted in a similar manner to those in Fig. 3.



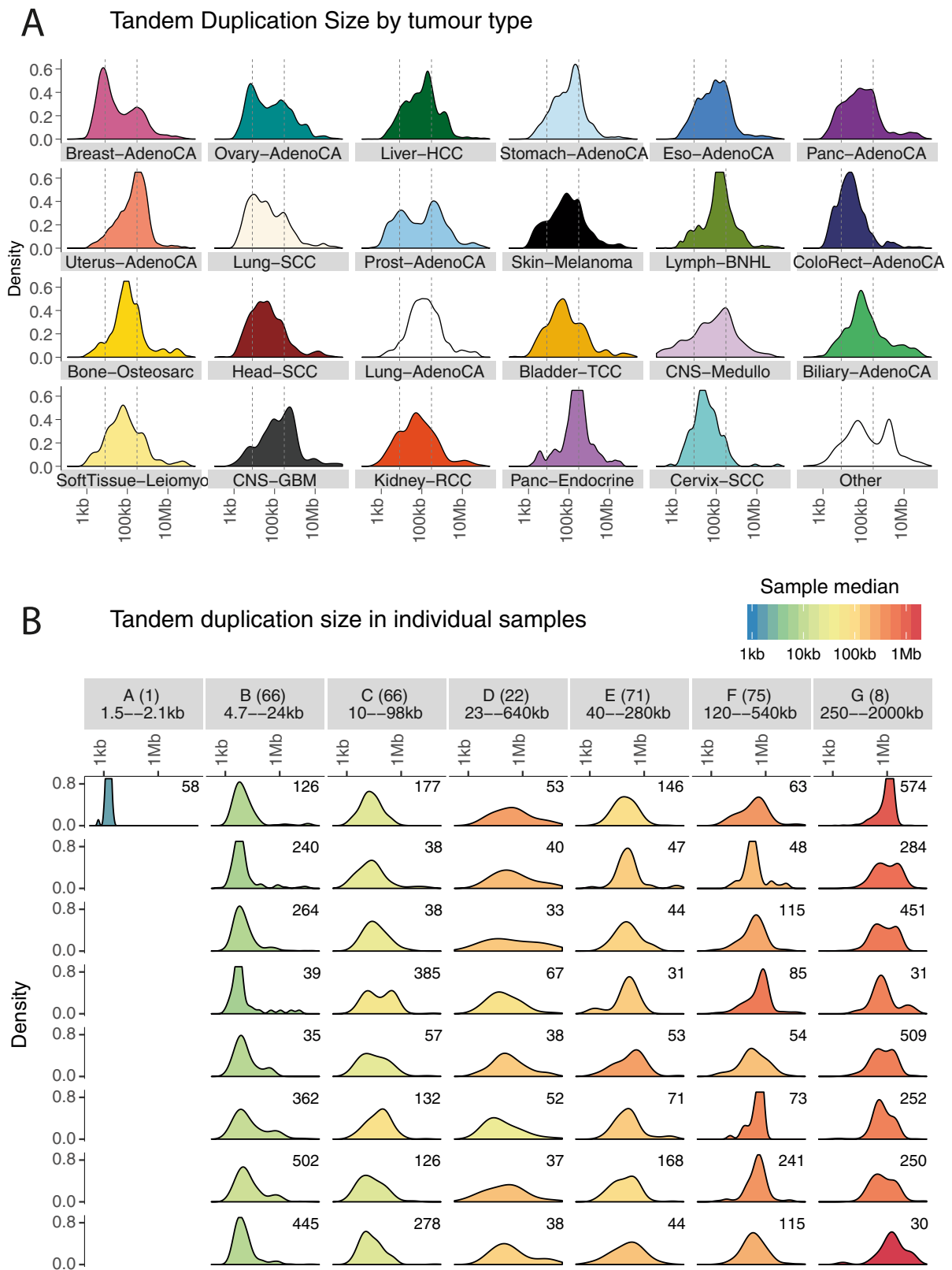
**Extended Data Fig. 4 | Templated insertion events that activate *TERT* in hepatocellular carcinoma.** **a**, The positions of all structural-variant breakpoints in the *TERT* region in the PCAWG cohort (including 50-kb flanks either side of *TERT*), coloured by classification and vertically spaced by the distance to the next breakpoint in the cohort. If the two sides of a breakpoint junction are contained within the plotting window, they are joined by a curved line. The number of samples with a breakpoint in the plotting window is annotated in the table in the top left. **b–d**, Examples of two cycles and a chain of

templated insertions that affect *TERT* in hepatocellular carcinomas. **e**, Expression levels of *TERT* in patients with hepatocellular carcinoma ( $n = 187$  patients), separated by whether *TERT* was wild type, had an activating promoter point mutation, structural variants in a templated insertion or other class. Individual patient data are shown as points. The box shows the median expression level as a thick black line, with the range of the box denoting the interquartile range. The whiskers show the range of data or 1.5× the interquartile range (whichever is lower).



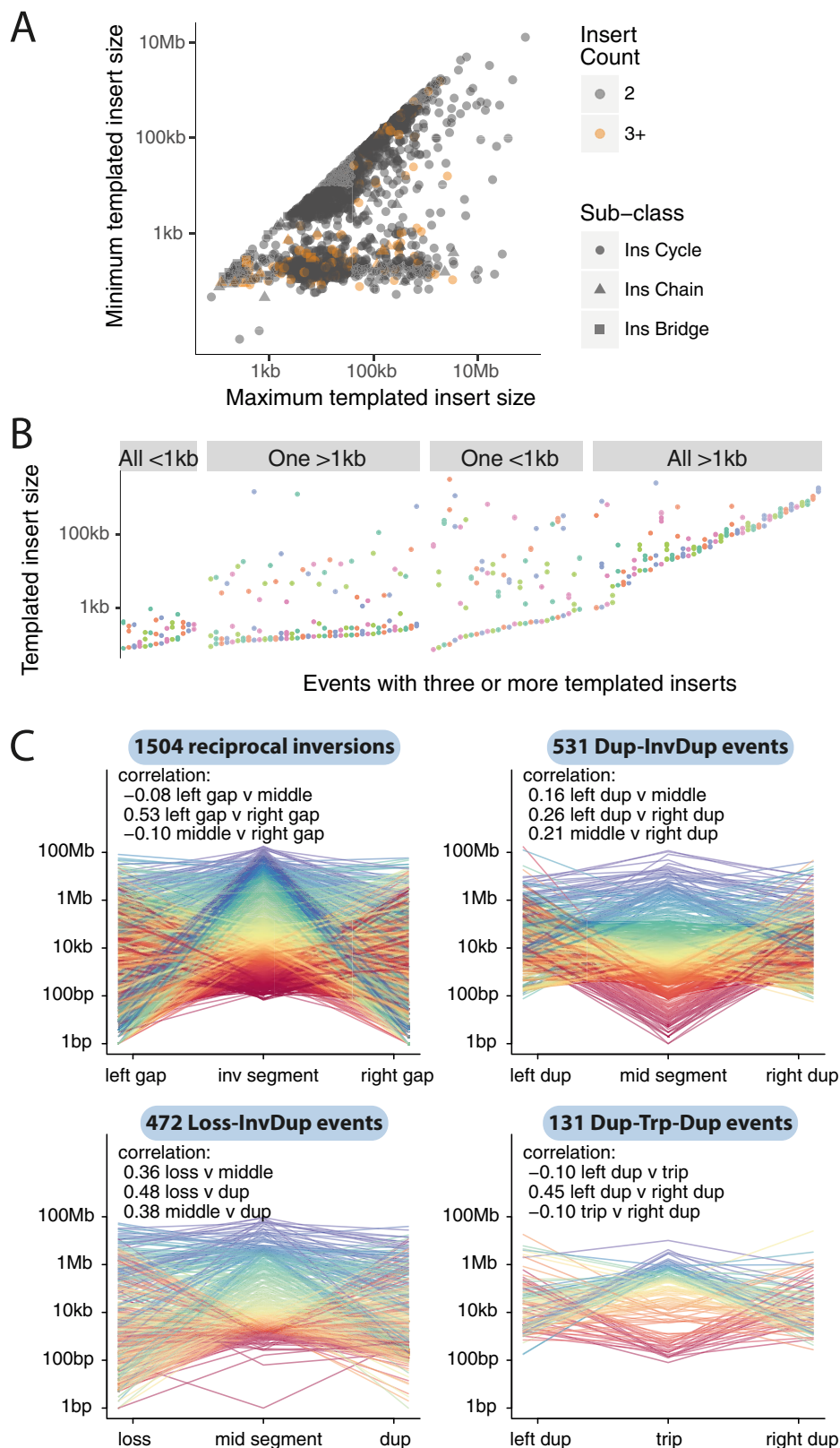
**Extended Data Fig. 5 | Templated insertion events inactivating *RB1* in breast and ovarian carcinomas.** **a**, The positions of all structural-variant breakpoints in the *RB1* region in the PCAWG cohort (including 50-kb flanks either side of *RB1*), coloured by classification and vertically spaced by the distance to the next breakpoint in the cohort. If the two sides of a breakpoint junction are

contained within the plotting window, they are joined by a curved line. The number of samples with a breakpoint in the plotting window is annotated in the table in the top left. **b-e**, Examples of three cycles and a bridge of templated insertions that affect *RB1* in breast and ovarian carcinomas.



**Extended Data Fig. 6 | Size distribution of tandem duplications.** **a**, Size distribution of tandem duplications per histology group. **b**, Samples with more than 20 tandem duplications were grouped using hierarchical clustering according to the within-patient distribution of tandem-duplication size. Seven

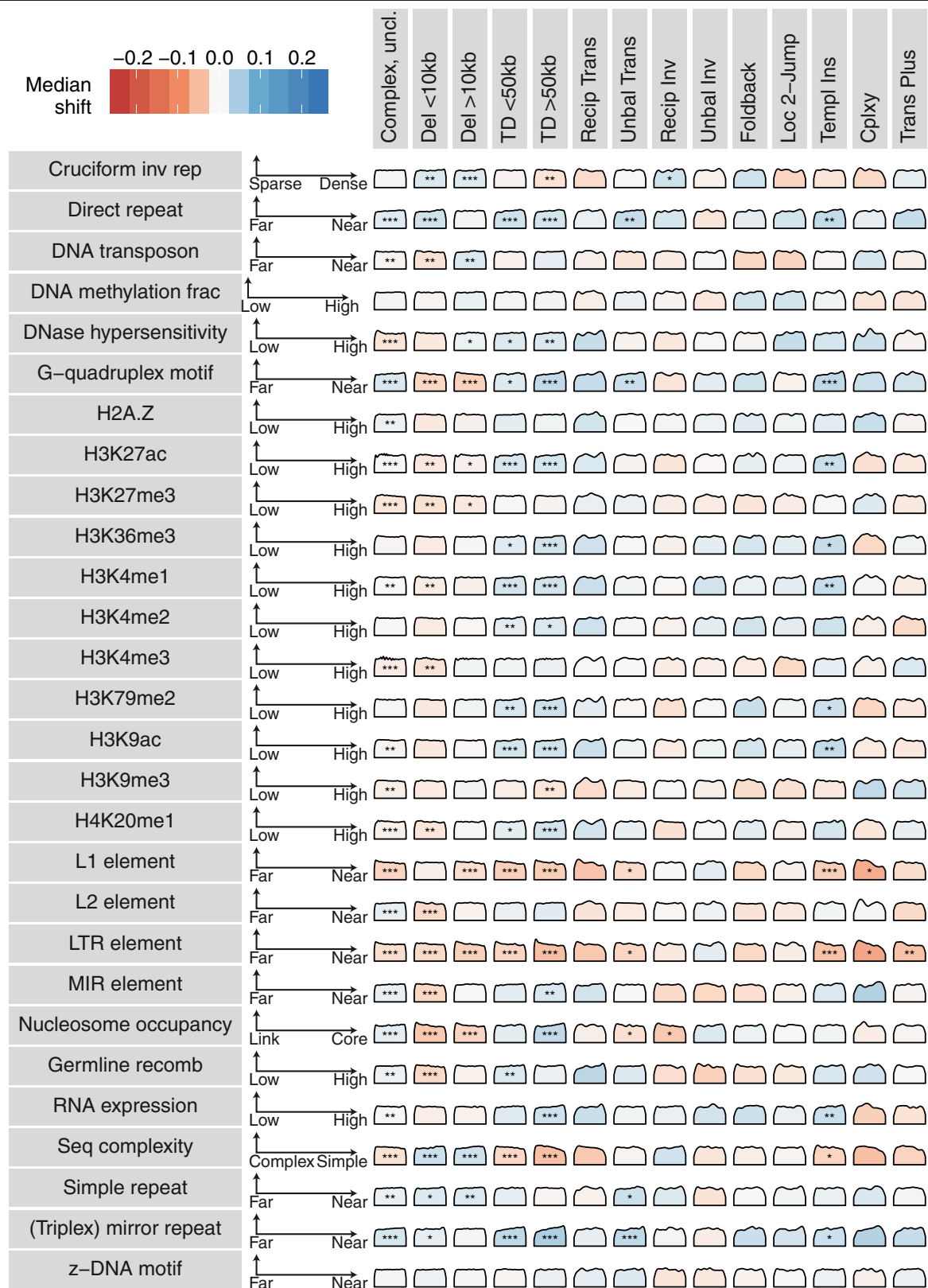
clusters emerged, with the size distribution of up to eight randomly chosen samples per cluster illustrated. The numbers in the top right of each panel denote the number of tandem duplications in that sample.



**Extended Data Fig. 7 | Size properties of clustered structural-variant classes.** **a**, Comparison of the minimum and maximum templated-insert size for multi-insert cycles, chains and bridges of templated insertions. **b**, All events with three or more templated inserts, grouped by combination of insert sizes. **c**, Correlations (Pearson's correlation coefficient) and raw sizes of individual

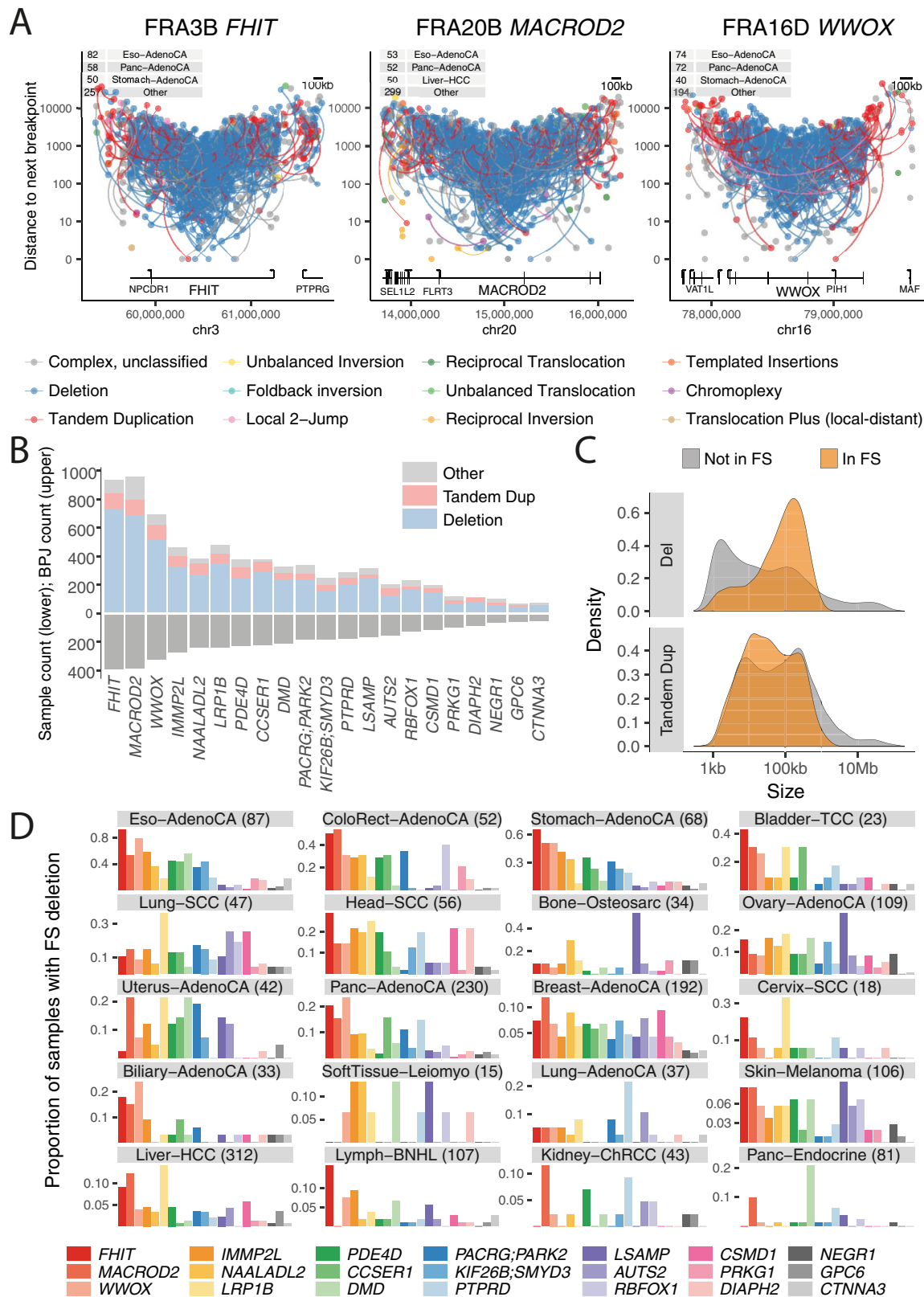
genomic segments for reciprocal inversions and local two-jumps. Each individual event is shown as a line that links the size of the individual segments in that event. The sample sizes for each event class are shown in the labels for each panel.





**Extended Data Fig. 8 | Relationship of an extended panel of genomic properties with structural-variant categories.** Associations between a subset of the genomic properties (rows) and classes of structural variant (columns). Each density curve represents the quantile distribution of the genomic property values at observed breakpoints, compared to random genome positions. Asterisks indicate significant departures from uniform quantiles after multiple hypothesis correction by the Benjamini-Yekutieli method on a

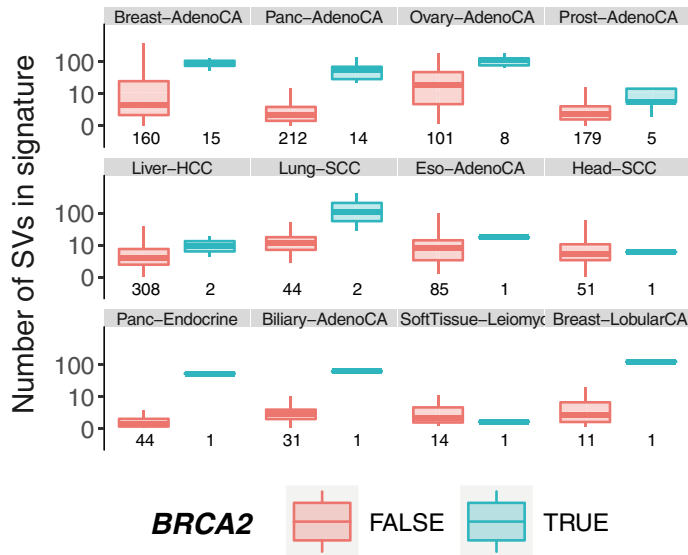
one-sided Kolmogorov-Smirnov test, based on a sample size of 2,559 genomes containing structural variants: \*false-discovery rate < 0.01, \*\*false-discovery rate < 0.001, \*\*\*false-discovery rate <  $10^{-6}$ . Cells with significant property associations are shaded by the magnitude of the shift of the median observed quantile above (blue) or below (red) 0.5. The interpretation of each property from left to right is indicated by the axes to the right of the property label.



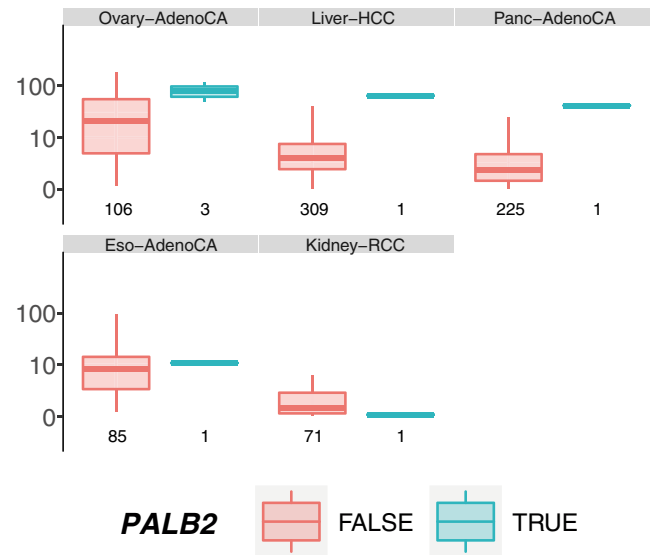
**Extended Data Fig. 9 | Properties of structural variants at chromosomal fragile sites.** **a**, Structural-variant breakpoints in the most affected fragile sites: *FHIT*, *MACROD2* and *WWOX*. These are coloured by classification and vertically spaced by the distance to the next breakpoint in the cohort. If the two sides of a breakpoint junction are contained within the plotting window, they are joined by a curved line. The number of samples with a breakpoint in the plotting window is annotated in the tables at the top left. **b**, Number of

deletions and tandem duplications (top) and number of affected samples (bottom) for the 18 fragile sites considered in this analysis. **c**, Size distribution of deletions and tandem duplications in fragile sites (FS) compared to the rest of the genome. **d**, Fragile-site preference for 20 cancer histology groups as indicated by the proportion of samples that contains a deletion in each of the 18 fragile sites considered here. The number of samples is indicated in parentheses.

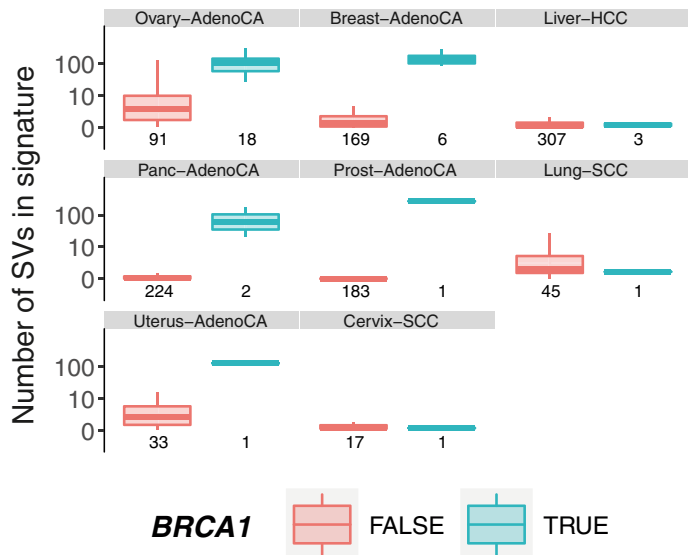
A

Small del SV signature vs *BRCA2* mutation

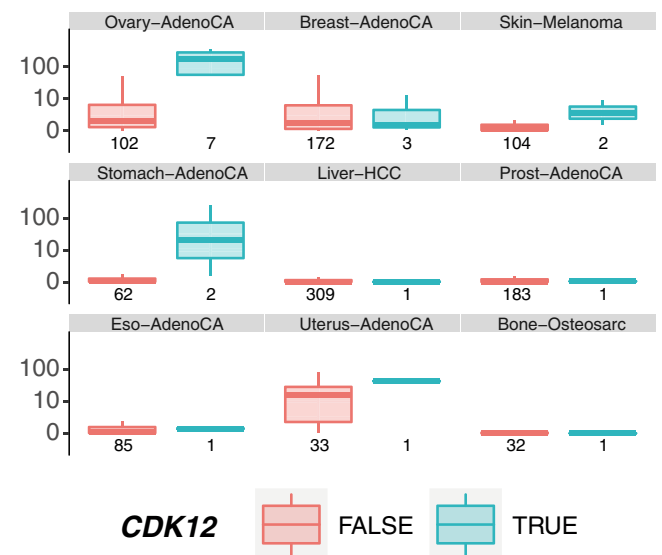
B

Small del SV signature vs *PALB2* mutation

C

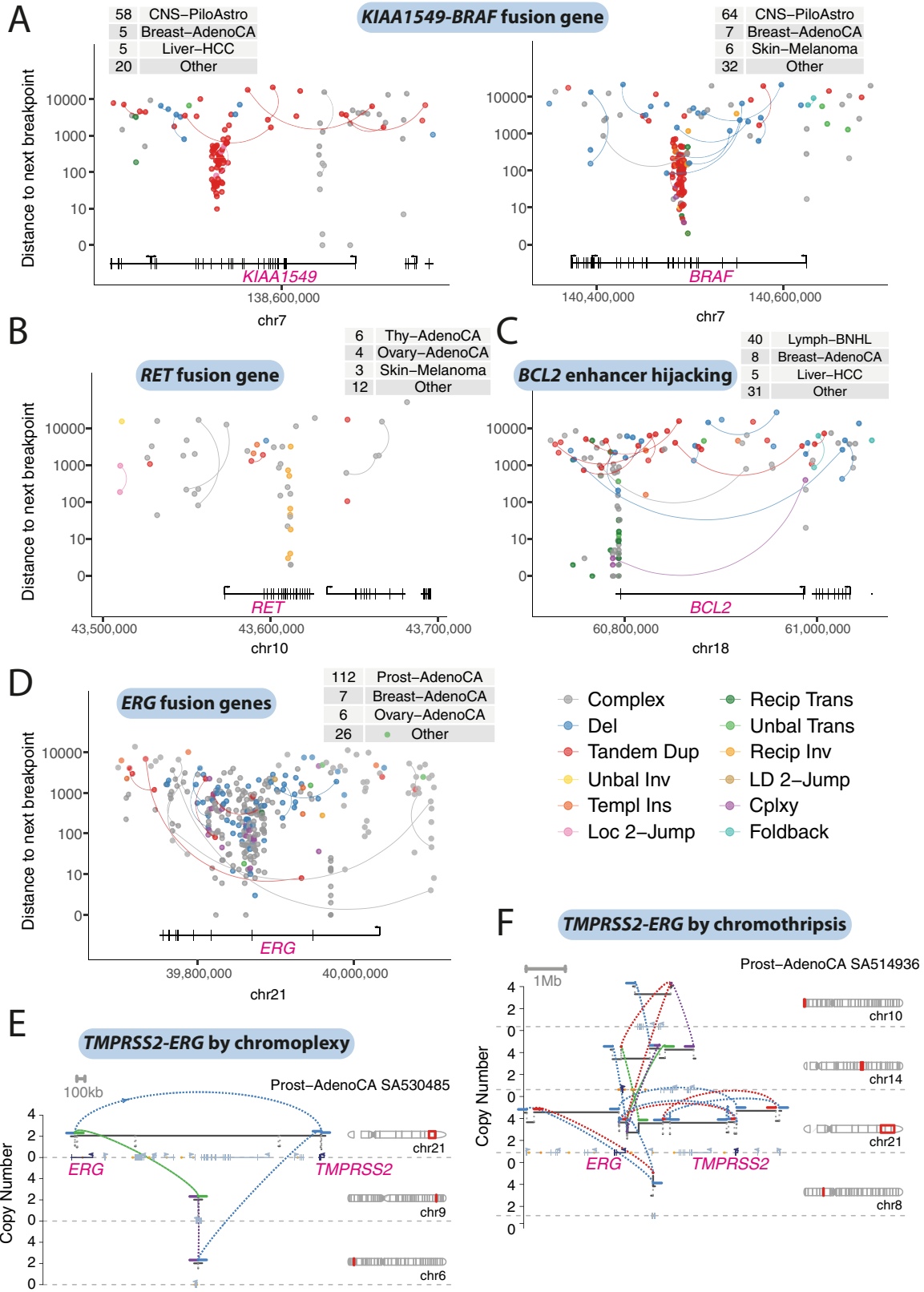
Early, small TD SV signature vs *BRCA1* mutation

D

Large TD SV signature vs *CDK12* mutation

**Extended Data Fig. 10 | Consistency of associations between signatures and mutations in DNA-repair genes.** **a**, Box-and-whisker plots showing the number of structural variants attributed to the small-deletion signature in different types of tumour, split by *BRCA2* status (*BRCA2* wild type in orange; *BRCA2* mutant in cyan). The box denotes the interquartile range, with the median marked as a horizontal line. The whiskers extend as far as the range or  $1.5 \times$  the interquartile range, whichever is lower. Outlier patients are shown as points. There is an increase in events attributed to the small-deletion signature when *BRCA2* is mutated, across multiple types of tumour (breast, pancreatic,

ovarian, prostate, lung squamous and so on). **b**, Box-and-whisker plots as for **a**, showing the number of structural variants attributed to the small-deletion signature in different types of tumour, split by *PALB2* status. **c**, Box-and-whisker plots as for **a**, showing the number of structural variants attributed to the early-replicating, small-tandem-duplication signature in different types of tumour, split by *BRCA1* status. **d**, Box-and-whisker plots as for **a**, showing the number of structural variants attributed to the large-tandem-duplication signature in different types of tumour, split by *CDK12* status.

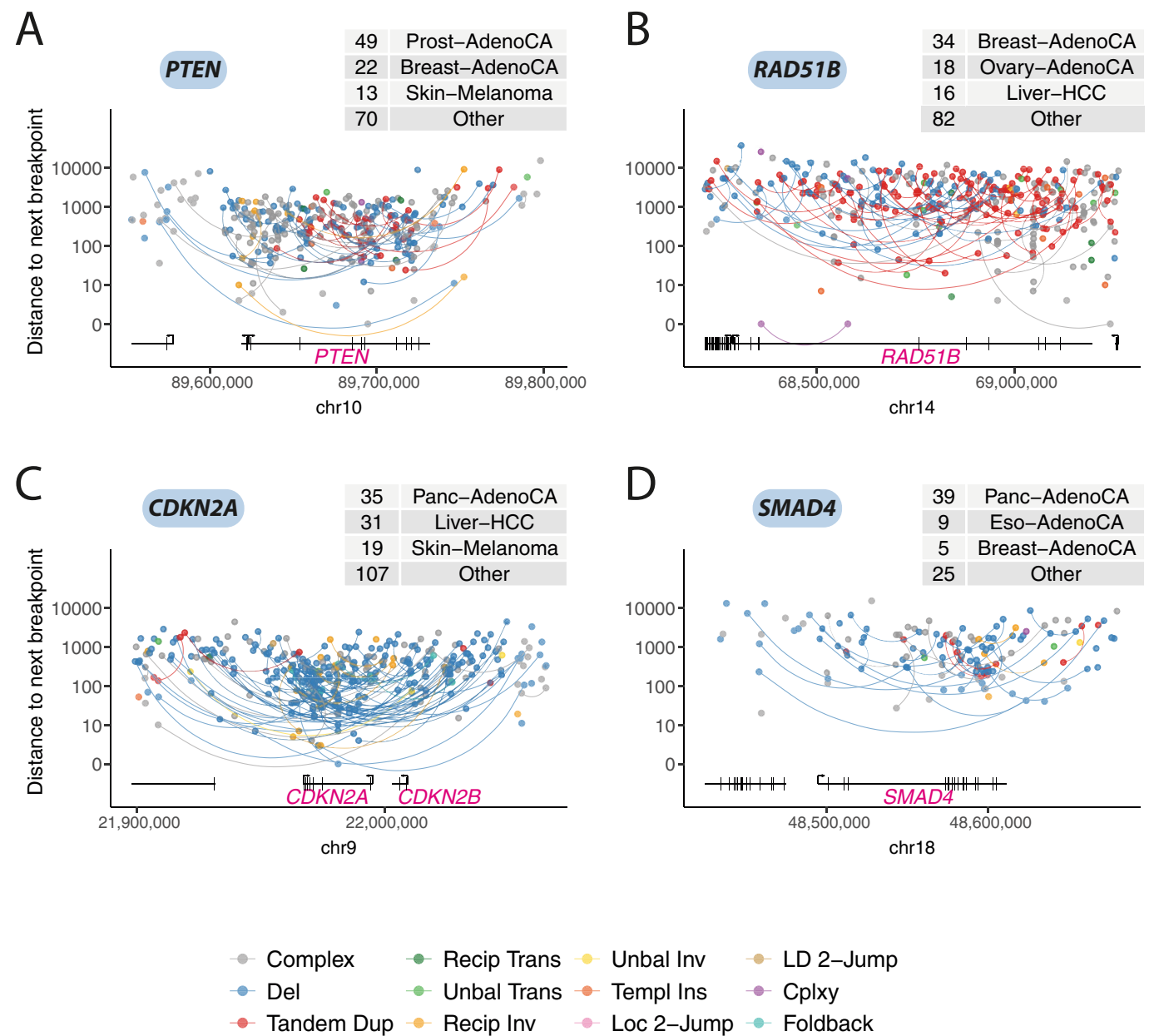


Extended Data Fig. 11 | See next page for caption.

**Extended Data Fig. 11 | Patterns of structural variants causing fusion genes and enhancer hijacking.** **a**, Rainfall plot of structural-variant breakpoints in the genes *KIAA1549* and *BRAF*, commonly fused together through a tandem duplication in pilocytic astrocytomas. Structural variants are coloured by classification and arranged vertically by the distance to the next breakpoint in the cohort. If the two sides of a breakpoint junction are contained within the plotting window, they are joined by a curved line. The number of samples with a breakpoint in the plotting window is annotated in the table at the top of each panel. **b**, Rainfall plot of structural-variant breakpoints that affect *RET*, commonly fused to *CCDC6* by inversion in papillary thyroid cancer. **c**, Rainfall

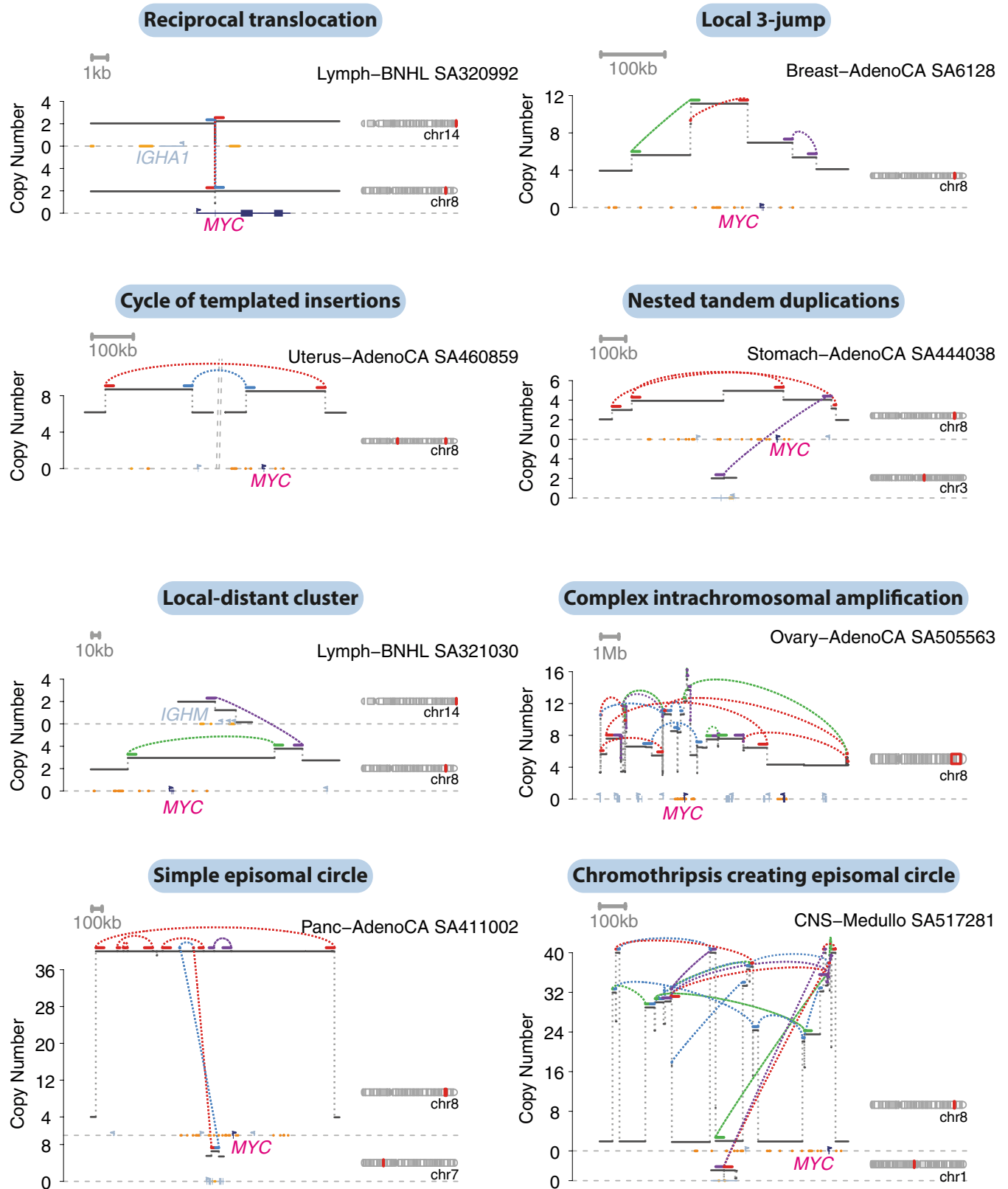
plot of structural-variant breakpoints that affect *BCL2*, commonly hijacked to the *IGH* immunoglobulin locus by translocations in B cell lymphomas. **d**, Rainfall plot of structural-variant breakpoints that affect *ERG*, commonly fused with *TMPRSS2* by deletion or more-complex events in prostate adenocarcinoma. **e**, Example of a *TMPRSS2-ERG* fusion gene in a prostate adenocarcinoma created by a chromoplexy cycle. The estimated copy-number profile is shown as black horizontal segments, with structural variants shown as dotted arcs linking the edges of two copy-number segments. **f**, Example of a *TMPRSS2-ERG* fusion gene in a prostate adenocarcinoma created by chromothripsis.





**Extended Data Fig. 12 | Patterns of structural variants that affect selected tumour-suppressor genes.** **a**, Rainfall plot of structural-variant breakpoints in the gene *PTEN*, commonly inactivated in breast and ovarian adenocarcinomas, in which tandem-duplication signatures are frequent. Structural variants are coloured by classification and arranged vertically by the distance to the next breakpoint in the cohort. If the two sides of a breakpoint junction are contained within the plotting window, they are joined by a curved line. The number of samples with a breakpoint in the plotting window is annotated in the table at

the top of each panel. **b**, Rainfall plot of structural-variant breakpoints that affect *RAD51B*, commonly inactivated in breast and ovarian adenocarcinomas. **c**, Rainfall plot of structural-variant breakpoints that affect *CDKN2A*, commonly inactivated in tumours of the gastrointestinal tract, in which deletion signatures are common. **d**, Rainfall plot of structural-variant breakpoints that affect *SMAD4*, commonly inactivated in tumours of the gastrointestinal tract.



**Extended Data Fig. 13 | Examples of structural variants increasing the copy number of *MYC*.** The estimated copy-number profile is shown as black horizontal segments, with structural variants shown as dotted arcs linking the edges of two copy-number segments.

Extended Data Table 1 | Glossary of key terms

Term	Description
Structural variant (SV)	Juxtaposition of non-contiguous chromosomal segments through a process of genomic rearrangement.
Breakpoint	The chromosomal position at which a DNA break is made. Each SV consists of a junction between two breakpoints in different regions of the genome.
Copy number alteration (CNA)	Change in the number of copies of a given chromosomal segment from that expected.
Reciprocal or balanced SV	A pair of SVs in which both sides of a single dsDNA break are rescued in the rearrangement. Typically used to describe some inversions and some translocations.
Unbalanced SV	An SV (usually inversion or translocation) in which only one side of the dsDNA break is rescued, thereby generating a copy number alteration across the breakpoint.
Cluster of SVs	A set of SVs that are closer together in genomic space than expected by chance. Typically, such clustering implies a shared mechanistic basis for the SV generation.
Derivative chromosome	A chromosome that carries one or more SVs.
Phased SVs	Set of SVs and copy number alterations in a cluster carried on a single derivative chromosome.
Chromosomal segment	A contiguous stretch of DNA that is of constant copy number, used to denote the regions of chromosome between SVs.
Template	A region of chromosomal DNA that is copied and inserted elsewhere in the genome.
<b>SV class</b>	A type of structural variant, such as deletion, tandem duplication or translocation.
Deletion	Loss of a segment of chromosome from the genome spanned by a junction between the two breakpoints either side.
Tandem duplication	Extra copy of a segment of chromosome in which the duplicated region is inserted immediately adjacent to the original template in the same orientation.
Reciprocal inversion	A segment of chromosomal DNA inserted into its original position, but in the opposite orientation.
Fold-back inversion	An inverted rearrangement between two breakpoints typically <20kb apart on the chromosome, with associated copy number change. Often a sign of breakage-fusion-bridge cycles.
Translocation	Breakpoint junction between two different chromosomes, either reciprocal or unbalanced.
Breakage-fusion-bridge cycle	SV mechanism in which a naked DNA end ( <i>breakage</i> ) is copied to its sister chromatid during S phase, with the two ends undergoing <i>fusion</i> (by fold-back inversion). At anaphase, the resulting dicentric chromosome is stretched between the two daughter cells ( <i>bridge</i> ), leading to further DNA breakage and potentially further cycles.
Chromoplexy	A set of >2 reciprocal SVs in which the chromosomal ends either side of each breakpoint are shuffled such that every end is rescued in a rearrangement junction.
Chromothripsis	A cluster of many SVs (10s to 100s) in one or a few chromosomes, occurring in a single catastrophic event, with oscillating copy number profile and rearrangement junctions of all four possible orientations.
Local <i>n</i> -jump	A cluster of <i>n</i> SVs in a single genomic region, typically phased to a single derivative chromosome, exhibiting some copy number gains and junctions with inverted and non-inverted orientation.
Cycle, chain or bridge of templated insertions	Copies of one or more genomic templates drawn from across the genome, strung together in a contiguous string and inserted into a single derivative chromosome. A <i>chain</i> of templated insertions does not return to the original chromosome, leading to an unbalanced translocation. A <i>cycle</i> has a duplication on the host chromosome, while a <i>bridge</i> inserts the template copies into a deletion on the host chromosome.
Local-distant cluster	A cluster of SVs that has both local rearrangements and rearrangements to other parts of the genome.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a                                 | Confirmed  |
|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted<br><i>Give P values as exact values whenever suitable.</i>                     |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated   |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

Workflow Step Algorithm Version Dockstore Package\*  
 WGS Alignment BWA-MEM v0.7.8-r455 <https://goo.gl/oqp4Xd>  
 EMBL/DKFZ SV caller DELLY v0.6.6 <https://goo.gl/Y46MCo>  
 EMBL/DKFZ SCNA caller ACESeq v1.0.189 <https://goo.gl/4zoV42>  
 EMBL/DKFZ SNV caller DKFZ somatic SNV workflow 1.0.132-1  
 EMBL/DKFZ indel caller Platypus v0.7.4  
 Sanger SCNA caller ascatNgs v1.5.2 <https://goo.gl/9DSrbA>  
 Sanger SV caller BRASS v4.012  
 grass v1.1.6  
 Sanger SNV caller CaVEMan v1.50  
 Sanger indel caller Pindel v1.5.7  
 Broad SCNA caller ABSOLUTE/JaBbA v1.5/? <https://goo.gl/YkdtDt>  
 Broad SV caller SvABA/dRanger/BreakPointer 2015-05-20/2016-03-13/2015-12-22  
 Broad SNV caller MuTect v1.1.4  
 Broad indel caller SvABA 2015-05-20  
 MuSE SNV caller MuSE v1.0rc <https://goo.gl/5SR4bF>  
 SMuFIN indel caller SMuFIN 2014-10-26 <https://goo.gl/EuUP5k>  
 Oxidative artefact filter OxoG 2016-4-28 <https://goo.gl/cUKP9K>  
 SNV/Indel annotation VAGrENT v2.1.2 <https://goo.gl/9DSrbA>  
 ANNOVAR v2014Nov12 <https://goo.gl/4zoV42>  
 miniBAM generation VariantBAM v2017Dec12 <https://goo.gl/S8h8e5>  
 SNV/Indel merging and consensus generation SNV-MERGE v2017May26 <https://goo.gl/TETSB8>  
 SV merging and consensus generation SV-MERGE v2017Dec12 <https://goo.gl/A9CEup>  
 Strand bias filter DKFZ Strand Bias Filter v2016Dec15 <https://goo.gl/8jXrvZ>

## Data analysis

We have described the algorithms in detail throughout the manuscript. Signatures analysis used the published NMF algorithm (PMID: 23318258) and an unpublished Bayesian method, released on github (<https://github.com/nicolaroberts/hdp>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The PCAWG-generated alignments, variant calls, annotations and derived data sets are available for general research use for browsing and download at <http://dcc.icgc.org/pcawg/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identifying information, such as germline alleles and underlying read data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the data set, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

## Sample size

We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014.

No formal power calculations were performed to decide sample size for PCAWG - we aggregated all genomes available at the time.

## Data exclusions

After quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine). Data exclusion criteria were pre-established.

## Replication

In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. To assess accuracy of SV calls, we therefore used the property that an SV must either generate a copy number change or be balanced, whereas artefactual calls will not respect this property. For individual SV callers, we estimated the true positive rate to be in the range 80-95% for samples in the pilot-63 dataset.

## Randomization

Not applicable - this was a descriptive study.

## Blinding

Not applicable - this was a descriptive study.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.



## Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

## Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

### Population characteristics

Patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-O-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.

### Recruitment

Patients were recruited by the participating centres following local protocols. Samples obtained had to meet criteria on amount of tumour DNA available, meaning that the cohort is potentially somewhat biased towards larger tumours. Otherwise, we anticipate no major recruitment biases.

### Ethics oversight

The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# The evolutionary history of 2,658 cancers

<https://doi.org/10.1038/s41586-019-1907-7>

Received: 11 August 2017

Accepted: 18 November 2019

Published online: 5 February 2020

Open access

Moritz Gerstung<sup>1,2,3,4,5\*</sup>, Clemency Jolly<sup>4,40</sup>, Ignaty Leshchiner<sup>5,40</sup>, Stefan C. Dentre<sup>3,4,6,40</sup>, Santiago Gonzalez<sup>1,40</sup>, Daniel Rosebrock<sup>5</sup>, Thomas J. Mitchell<sup>3,7</sup>, Yulia Rubanova<sup>8,9</sup>, Pavana Anur<sup>10</sup>, Kaixian Yu<sup>11</sup>, Maxime Tarabichi<sup>3,4</sup>, Amit Deshwar<sup>8,9</sup>, Jeff Wintersinger<sup>8,9</sup>, Kortine Kleinheinz<sup>12,13</sup>, Ignacio Vázquez-García<sup>3,7</sup>, Kerstin Haase<sup>4</sup>, Lara Jerman<sup>1,14</sup>, Subhajit Sengupta<sup>15</sup>, Geoff Macintyre<sup>16</sup>, Salem Malikic<sup>17,18</sup>, Nilgun Donmez<sup>17,18</sup>, Dimitri G. Livitz<sup>5</sup>, Marek Cmero<sup>19,20</sup>, Jonas Demeulemeester<sup>4,21</sup>, Steven Schumacher<sup>5</sup>, Yu Fan<sup>11</sup>, Xiaotong Yao<sup>22,23</sup>, Juhee Lee<sup>24</sup>, Matthias Schlesner<sup>12</sup>, Paul C. Boutros<sup>8,25,26</sup>, David D. Bowtell<sup>27</sup>, Hongtu Zhu<sup>11</sup>, Gad Getz<sup>5,28,29,30</sup>, Marcin Imielinski<sup>22,23</sup>, Rameen Beroukhi<sup>5,31</sup>, S. Cenik Sahinalp<sup>18,32</sup>, Yuan Ji<sup>15,33</sup>, Martin Peifer<sup>34</sup>, Florian Markowitz<sup>16</sup>, Ville Mustonen<sup>35</sup>, Ke Yuan<sup>16,36</sup>, Wenyi Wang<sup>11</sup>, Quaid D. Morris<sup>8,9</sup>, PCAWG Evolution & Heterogeneity Working Group<sup>37</sup>, Paul T. Spellman<sup>10,41</sup>, David C. Wedge<sup>6,38,41</sup>, Peter Van Loo<sup>4,21,41\*</sup> & PCAWG Consortium<sup>39</sup>

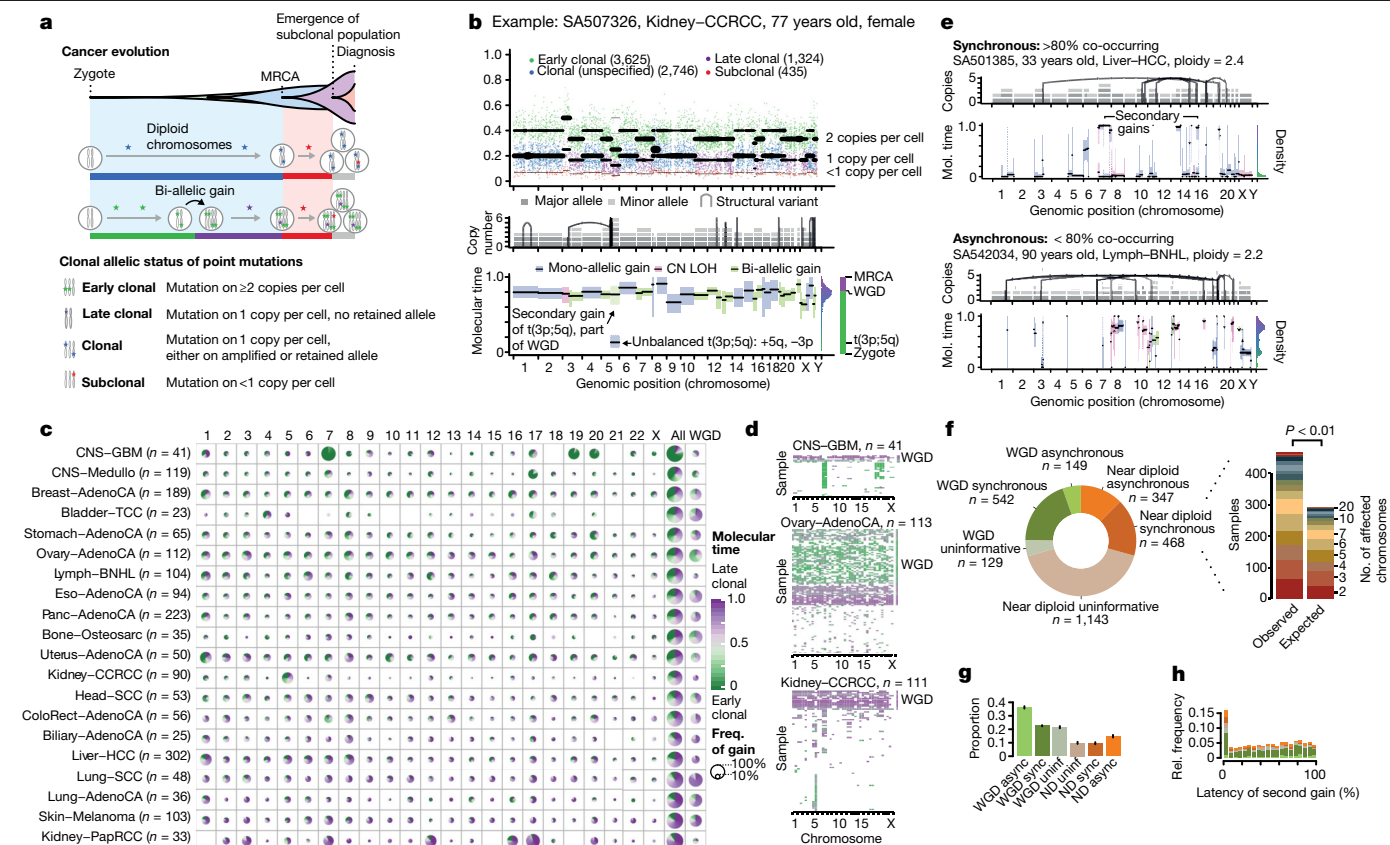
Cancer develops through a process of somatic evolution<sup>1,2</sup>. Sequencing data from a single biopsy represent a snapshot of this process that can reveal the timing of specific genomic aberrations and the changing influence of mutational processes<sup>3</sup>. Here, by whole-genome sequencing analysis of 2,658 cancers as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)<sup>4</sup>, we reconstruct the life history and evolution of mutational processes and driver mutation sequences of 38 types of cancer. Early oncogenesis is characterized by mutations in a constrained set of driver genes, and specific copy number gains, such as trisomy 7 in glioblastoma and isochromosome 17q in medulloblastoma. The mutational spectrum changes significantly throughout tumour evolution in 40% of samples. A nearly fourfold diversification of driver genes and increased genomic instability are features of later stages. Copy number alterations often occur in mitotic crises, and lead to simultaneous gains of chromosomal segments. Timing analyses suggest that driver mutations often precede diagnosis by many years, if not decades. Together, these results determine the evolutionary trajectories of cancer, and highlight opportunities for early cancer detection.

Similar to the evolution in species, the approximately  $10^{14}$  cells in the human body are subject to the forces of mutation and selection<sup>1</sup>. This process of somatic evolution begins in the zygote and only comes to rest at death, as cells are constantly exposed to mutagenic stresses, introducing 1–10 mutations per cell division<sup>2</sup>. These mutagenic forces lead to a gradual accumulation of point mutations throughout life, observed in a range of healthy tissues<sup>5–11</sup> and cancers<sup>12</sup>. Although these mutations are predominantly selectively neutral passenger mutations, some are proliferatively advantageous driver mutations<sup>13</sup>. The types of mutation in cancer genomes are well studied, but little is known

about the times when these lesions arise during somatic evolution and where the boundary between normal evolution and cancer progression should be drawn.

Sequencing of bulk tumour samples enables partial reconstruction of the evolutionary history of individual tumours, based on the catalogue of somatic mutations they have accumulated<sup>3,14,15</sup>. These inferences include timing of chromosomal gains during early somatic evolution<sup>16</sup>, phylogenetic analysis of late cancer evolution using matched primary and metastatic tumour samples from individual patients<sup>17–20</sup>, and temporal ordering of driver mutations across many samples<sup>21,22</sup>.

<sup>1</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. <sup>2</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. <sup>3</sup>Wellcome Sanger Institute, Cambridge, UK. <sup>4</sup>The Francis Crick Institute, London, UK. <sup>5</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>6</sup>Big Data Institute, University of Oxford, Oxford, UK. <sup>7</sup>University of Cambridge, Cambridge, UK. <sup>8</sup>University of Toronto, Toronto, Ontario, Canada. <sup>9</sup>Vector Institute, Toronto, Ontario, Canada. <sup>10</sup>Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. <sup>11</sup>The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>12</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>13</sup>Heidelberg University, Heidelberg, Germany. <sup>14</sup>University of Ljubljana, Ljubljana, Slovenia. <sup>15</sup>NorthShore University HealthSystem, Evanston, IL, USA. <sup>16</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>17</sup>Simon Fraser University, Burnaby, British Columbia, Canada. <sup>18</sup>Vancouver Prostate Centre, Vancouver, British Columbia, Canada. <sup>19</sup>University of Melbourne, Melbourne, Victoria, Australia. <sup>20</sup>Walter and Eliza Hall Institute, Melbourne, Victoria, Australia. <sup>21</sup>University of Leuven, Leuven, Belgium. <sup>22</sup>Weill Cornell Medicine, New York, NY, USA. <sup>23</sup>New York Genome Center, New York, NY, USA. <sup>24</sup>University of California Santa Cruz, Santa Cruz, CA, USA. <sup>25</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>26</sup>University of California, Los Angeles, CA, USA. <sup>27</sup>Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. <sup>28</sup>Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA, USA. <sup>29</sup>Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. <sup>30</sup>Harvard Medical School, Boston, MA, USA. <sup>31</sup>Dana-Farber Cancer Institute, Boston, MA, USA. <sup>32</sup>Indiana University, Bloomington, IN, USA. <sup>33</sup>The University of Chicago, Chicago, IL, USA. <sup>34</sup>University of Cologne, Cologne, Germany. <sup>35</sup>University of Helsinki, Helsinki, Finland. <sup>36</sup>University of Glasgow, Glasgow, UK. <sup>37</sup>A list of members and their affiliations appears at the end of the paper. <sup>38</sup>Oxford NIHR Biomedical Research Centre, Oxford, UK. <sup>39</sup>A list of members and their affiliations appears in the Supplementary Information. <sup>40</sup>These authors contributed equally: Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentre, Santiago Gonzalez. <sup>41</sup>These authors jointly supervised this work: Paul T. Spellman, David C. Wedge, Peter Van Loo. \*e-mail: moritz.gerstung@ebi.ac.uk; peter.vanloo@crick.ac.uk



**Fig. 1 | Timing clonal copy number gains using allele frequencies of point mutations.** **a**, Principles of timing mutations and copy number gains based on whole-genome sequencing. The number of sequencing reads reporting point mutations can be used to discriminate variants as early or late clonal (green or purple, respectively) in cases of specific copy number gains, as well as clonal (blue) or subclonal (red) in cases without. **b**, Annotated point mutations in one sample based on VAF (top), copy number (CN) state and structural variants (middle), and resulting timing estimates (bottom). LOH, loss of heterozygosity. **c**, Overview of the molecular timing distribution of copy number gains across cancer types. Pie charts depict the distribution of the inferred mutation time for a given copy number gain in a cancer type. Green denotes early clonal gains, with a gradient to purple for late gains. The size of each chart is proportional to the recurrence of this event. Abbreviations for each cancer type are defined

in Supplementary Table 1. **d**, Heat maps representing molecular timing estimates of gains on different chromosome arms (x axis) for individual samples (y axis) for selected tumour types. **e**, Temporal patterns of two near-diploid cases illustrating synchronous gains (top) and asynchronous gains (bottom). **f**, Left, distribution of synchronous and asynchronous gain patterns across samples, split by WGD status. Uninformative samples have too few or too small gains for accurate timing. Right, the enrichment of synchronous gains in near-diploid samples is shown by systematic permutation tests. **g**, Proportion of copy number segments ( $n = 90,387$ ) with secondary gains. Error bars denote 95% credible intervals. ND, near diploid. **h**, Distribution of the relative latency of  $n = 824$  secondary gains with available timing information, scaled to the time after the first gain and aggregated per chromosome.

The PCAWG Consortium has aggregated whole-genome sequencing data from 2,658 cancers<sup>4</sup>, generated by the ICGC and TCGA, and produced high-accuracy somatic variant calls, driver mutations, and mutational signatures<sup>4,23,24</sup> (Methods and Supplementary Information).

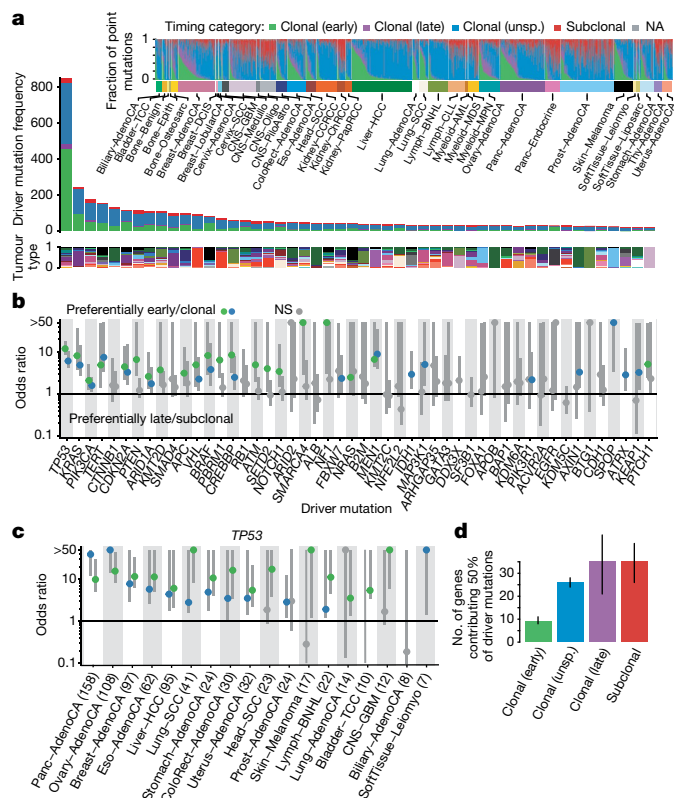
Here, we leverage the PCAWG dataset to characterize the evolutionary history of 2,778 cancer samples from 2,658 unique donors across 38 cancer types. We infer timing and patterns of chromosomal evolution and learn typical sequences of mutations across samples of each cancer type. We then define broad periods of tumour evolution and examine how drivers and mutational signatures vary between these epochs. Using clock-like mutational processes, we map mutation timing estimates into approximate real time. Combined, these analyses allow us to sketch out the typical evolutionary trajectories of cancer, and map them in real time relative to the point of diagnosis.

## Reconstructing the life history of tumours

The genome of a cancer cell is shaped by the cumulative somatic aberrations that have arisen during its evolutionary past, and part of this history can be reconstructed from whole-genome sequencing data<sup>3</sup> (Fig. 1a). Initially, each point mutation occurs on a single chromosome

in a single cell, which gives rise to a lineage of cells bearing the same mutation. If that chromosomal locus is subsequently duplicated, any point mutation on this allele preceding the gain will subsequently be present on the two resulting allelic copies, unlike mutations succeeding the gain, or mutations on the other allele. As sequencing data enable the measurement of the number of allelic copies, one can define categories of early and late clonal variants, preceding or succeeding copy number gains, as well as unspecified clonal variants, which are common to all cancer cells, but cannot be timed further. Lastly, we identify subclonal mutations, which are present in only a subset of cells and have occurred after the most recent common ancestor (MRCA) of all cancer cells in the tumour sample (Supplementary Information).

The ratio of duplicated to non-duplicated mutations within a gained region can be used to estimate the time point when the gain happened during clonal evolution, referred to here as molecular time, which measures the time of occurrence relative to the total number of (clonal) mutations. For example, there would be few, if any, co-amplified early clonal mutations if the gain had occurred right after fertilization, whereas a gain that happened towards the end of clonal tumour evolution would contain many duplicated mutations<sup>14</sup> (Fig. 1a, Methods).



**Fig. 2 | Timing of point mutations shows that recurrent driver gene mutations occur early.** **a**, Top, distribution of point mutations over different mutation periods in  $n = 2,778$  samples. Middle, timing distribution of driver mutations in the 50 most recurrent lesions across  $n = 2,583$  white listed samples from unique donors. Bottom, distribution of driver mutations across cancer types; colour as defined in the inset. **b**, Relative timing of the 50 most recurrent driver lesions, calculated as the odds ratio of early versus late clonal driver mutations versus background, or clonal versus subclonal. Error bars denote 95% confidence intervals derived from bootstrap resampling. Odds ratios overlapping 1 in less than 5% of bootstrap samples are considered significant (coloured). The underlying number of samples with a given mutation is shown in **a**. **c**, Relative timing of *TP53* mutations across cancer types, as in **b**. The number of samples is defined in the x-axis labels. **d**, Estimated number of unique lesions (genes) contributing 50% of all driver mutations in different timing epochs across  $n = 2,583$  unique samples, containing  $n = 5,756$  driver mutations with available timing information. Error bars denote the range between 0 and 1 pseudocounts; bars denote the average of the two values. NA, not applicable; NS, not significant.

These analyses are illustrated in Fig. 1b. As expected, the variant allele frequencies (VAFs) of somatic point mutations cluster around the values imposed by the purity of the sample, local copy number configuration and identified subclonal populations. The depicted clear cell renal cell carcinoma has gained chromosome arm 5q at an early molecular time as part of an unbalanced translocation  $t(3p;5q)$ , which confirms the notion that this lesion often occurs in adolescence in this cancer type<sup>16</sup>. At a later time point, the sample underwent a whole genome duplication (WGD) event, duplicating all alleles, including the derivative chromosome, in a single event, as evidenced by the mutation time estimates of all copy number gains clustering around a single time point, independently of the exact copy number state.

### Timing patterns of copy number gains

To systematically examine the mutational timing of chromosomal gains throughout the evolution of tumours in the PCAWG dataset, we applied this analysis to the 2,116 samples with copy number gains suitable for

timing (Supplementary Information). We find that chromosomal gains occur across a wide range of molecular times (median molecular time 0.60, interquartile range (IQR) 0.10–0.87), with systematic differences between tumour types, whereas within tumour types, different chromosomes typically show similar distributions (Fig. 1c, Extended Data Figs. 1, 2, Supplementary Information). In glioblastoma and medulloblastoma, a substantial fraction of gains occurs early in molecular time. By contrast, in lung cancers, melanomas and papillary kidney cancers, gains arise towards the end of the molecular timescale. Most tumour types, including breast, ovarian and colorectal cancers, show relatively broad periods of chromosomal instability, indicating a very variable timing of gains across samples.

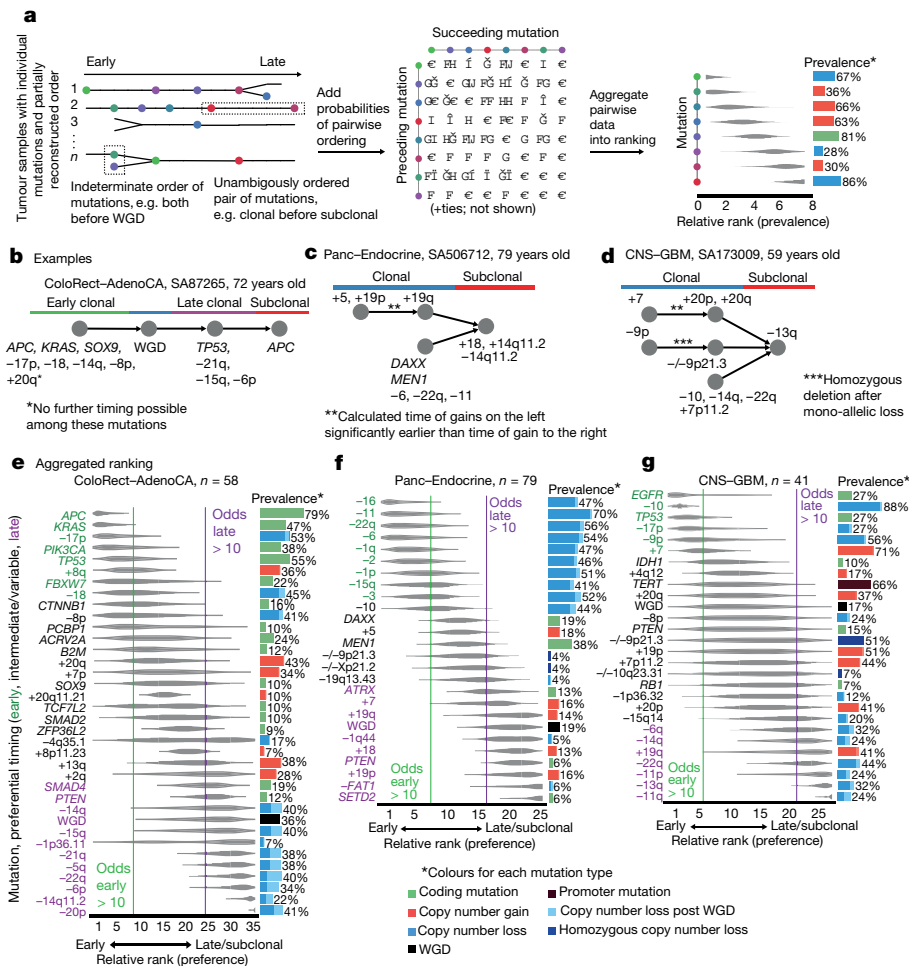
There are, however, certain tumour types with consistently early or late gains of specific chromosomal regions. Most pronounced is glioblastoma, in which 90% of tumours contain single copy gains of chromosome 7, 19 or 20 (Fig. 1c, d). Notably, these gains are consistently timed within the first 10% of molecular time, which suggests that they arise very early in a patient's lifetime. In the case of trisomy 7, typically less than 3 out of 600 single nucleotide variants (SNVs) on the whole chromosome precede the gain (Extended Data Fig. 3a, b). On the basis of a mutation rate of  $\mu = 4.8 \times 10^{-10}$  to  $3.0 \times 10^{-9}$  SNVs per base pair per division<sup>25</sup>, this indicates that the trisomy occurs within the first 6–39 cell divisions, suggesting a possible early developmental origin, in agreement with somatic mosaicism observed in the healthy brain<sup>26</sup>. Similarly, the duplications leading to isochromosome 17q in medulloblastoma are timed exceptionally early (Extended Data Fig. 3c, d).

Notably, we observed that gains in the same tumour often appear to occur at a similar molecular time, pointing towards punctuated bursts of copy number gains involving most gained segments (Fig. 1e). Although this is expected in tumours with WGD (Fig. 1b), it may seem surprising to observe synchronous gains in near-diploid tumours, particularly as only 6% of co-amplified chromosomal segments were linked by a direct inter-chromosomal structural variant. Still, synchronous gains are frequent, occurring in 57% (468 out of 815) of informative near-diploid tumours, 61% more frequently than expected by chance ( $P < 0.01$ , permutation test; Fig. 1f). Because most arm-level gains increment the allele-specific copy number by 1 (80–90%; Fig. 1g), it seems that these gains arise through mis-segregation of single copies during anaphase. This notion is further supported by the observation that in about 85% of segments with two gains of the same allele, the second gain appears with noticeable latency after the first (Fig. 1h). Therefore, the extensive chromosome-scale copy number aberrations observed in many cancer genomes are seemingly caused by a limited number of events—possibly by merotelic attachments of chromosomes to multipolar mitotic spindles<sup>27</sup>, or as a consequence of negative selection of individual aneuploidies<sup>28</sup>—offering an explanation for observations of punctuated evolution in breast and colorectal cancer<sup>29,30</sup>.

### Timing of point mutations in driver genes

As outlined above, point mutations (SNVs and insertions and deletions (indels)) can be qualitatively assigned to different epochs, allowing the timing of driver mutations. Out of the 47 million point mutations in 2,583 unique samples, 22% were early clonal, 7% late clonal, 53% unspecified clonal and 17% subclonal (Fig. 2a). Among a panel of 453 cancer driver genes, 5,913 oncogenic point mutations were identified<sup>4</sup>, of which 29% were early clonal, 5% late clonal, 56% unspecified clonal and 8% subclonal. It thus emerges that common drivers are enriched in the early clonal and unspecified clonal categories and depleted in the late clonal and subclonal ones, indicating a preferential early timing (Fig. 2b). For example, driver mutations in *TP53* and *KRAS* are 12 and 8 times enriched in early clonal stages, respectively. For *TP53*, this trend is independent of tumour type (Fig. 2c). Mutations in *PIK3CA* are two times more frequently clonal than expected, and non-coding changes near the *TERT* gene are three times more frequently early clonal.





**Fig. 3 | Aggregating single-sample ordering reveals typical timing of driver mutations.** **a**, Schematic representation of the ordering process. **b–d**, Examples of individual patient trajectories (partial ordering relationships), the constituent data for the ordering model process. **e–g**, Preferential ordering diagrams for colorectal adenocarcinoma (ColoRect-AdenoCA) (**e**), pancreatic

neuroendocrine cancer (Pano-Endocrine) (**f**) and glioblastoma (CNS-GBM) (**g**). Probability distributions show the uncertainty of timing for specific events in the cohort. Events with odds above 10 (either earlier or later) are highlighted. The prevalence of the event type in the cohort is displayed as a bar plot on the right.

Aggregating the clonal status of all driver point mutations over time reveals an increased diversity of driver genes mutated at later stages of tumour development: 50% of all early clonal driver mutations occur in just 9 genes, whereas 50% of late and subclonal mutations occur in approximately 35 different genes each, a nearly fourfold increase (Fig. 2d). Consistent with previous studies of individual tumour types<sup>31–34</sup>, these results suggest that, in general, the very early events in cancer evolution occur in a constrained set of common drivers, and a more diverse array of drivers is involved in late tumour development.

### Relative timing of somatic driver events

Although timing estimates of individual events reflect evolutionary periods that differ from one sample to another, they define in part the order in which driver mutations and copy number alterations have occurred in each sample (Fig. 3a–d). As confirmed by simulations, aggregating these orderings across samples defines a probabilistic ranking of lesions (Fig. 3a), recapitulating whether each mutation occurs preferentially early or late during tumour evolution (Extended Data Figs. 4, 5, Supplementary Information).

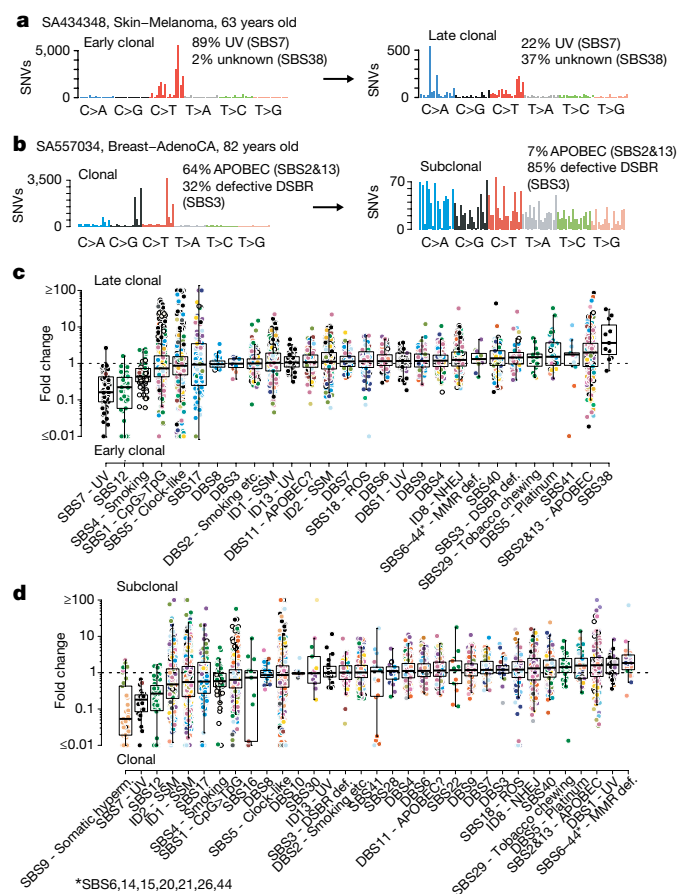
In colorectal adenocarcinoma, for example, we find APC mutations to have the highest odds of occurring early, followed by KRAS, loss of 17p and TP53, and SMAD4 (Fig. 3b, e). Whole-genome duplications

occur after tumours have accumulated several driver mutations, and many chromosomal gains and losses are typically late. These results are in agreement with the classical APC-KRAS-TP53 progression model of Fearon and Vogelstein<sup>35</sup>, but add considerable detail.

In many cancer types, the sequence of events during cancer progression has not previously been determined in detail. For example, in pancreatic neuroendocrine cancers, we find that many chromosomal losses, including those of chromosomes 2, 6, 11 and 16, are among the earliest events, followed by driver mutations in MEN1 and DAXX (Fig. 3c, f). WGD events occur later, after many of these tumours have reached a pseudo-haploid state due to widespread chromosomal losses. In glioblastoma, we find that the loss of chromosome 10, and driver mutations in TP53 and EGFR are very early, often preceding early gains of chromosomes 7, 19 and 20 (Fig. 3d, g). Mutations in the TERT promoter tend to occur at early to intermediate time points, whereas other driver mutations and copy number changes tend to be later events.

Across cancer types, we typically find TP53 mutations among the earliest events, as well as losses of chromosome 17 (Supplementary Information). WGD events usually have an intermediate ranking, and most copy number changes occur later. Losses typically precede gains, and consistent with the results above, common drivers typically occur before rare drivers.





**Fig. 4 | Dynamic mutational processes during early and late clonal tumour evolution.** **a**, Example of tumours with substantial changes between mutation spectra of early (left) and late (right) clonal time points. The attribution of mutations to the most characteristic signatures are shown. **b**, Example of clonal-to-subclonal mutation spectrum change. **c**, Fold changes between relative proportions of early and late clonal mutations attributed to individual mutational signatures. Points are coloured by tissue type. Data are shown for samples ( $n = 530$ ) with measurable changes in their overall mutation spectra and restricted to signatures active in at least 10 samples. Box plots demarcate the first and third quartiles of the distribution, with the median shown in the centre and whiskers covering data within  $1.5 \times$  the IQR from the box. **d**, Fold changes between clonal and subclonal periods in samples ( $n = 729$ ) with measurable changes in their mutation spectra, analogous to **c**.

### Timing of mutational signatures

The cancer genome is shaped by various mutational processes over its lifetime, stemming from exogenous and cell-intrinsic DNA damage, and error-prone DNA replication, leaving behind characteristic mutational spectra, termed mutational signatures<sup>24,36</sup>. Stratifying mutations by their clonal allelic status, we find evidence for a changing mutational spectrum between early and late clonal time points in 29% (530 out of 1,852) of informative samples ( $P < 0.05$ , Bonferroni-adjusted likelihood-ratio test), typically changing the spectrum by 19% (median absolute difference; range 4–66%) (Fig. 4a, b, Extended Data Fig. 6). Similarly, 30% of informative samples (729 out of 2,387) displayed changes of their mutation spectrum between the clonal and subclonal state, with median difference of 21% (range 3–72%). Combined, the mutation spectrum changes throughout tumour evolution in 40% of samples (1,069 out of 2,688).

To quantify whether the observed temporal changes can be attributed to known and suspected mutational processes, we decomposed the mutational spectra at each time point into a catalogue of

57 mutational signatures, including double base substitution and indel signatures<sup>24</sup> (Methods).

In general, these mutational signatures display a predominantly undirected temporal variability over several orders of magnitude (Fig. 4c, d, Extended Data Fig. 7). In addition, several signatures demonstrate distinct temporal trends. As one may expect, signatures of exogenous mutagens are predominantly active in the early clonal stages of tumorigenesis. These include tobacco smoking in lung adenocarcinoma (signature SBS4, median fold change 0.43, IQR 0.31–0.72), consistent with previous reports<sup>37,38</sup>, and ultraviolet light exposure in melanoma (SBS7; median fold change 0.16, IQR 0.09–0.43). Another strong decrease over time is found for a signature of unknown aetiology, SBS12, which acts mostly in liver cancers (median fold change 0.22, IQR 0.06–0.41). In chronic lymphoid leukaemia, there was a 20-fold relative decrease in mutations associated with somatic hypermutation (SBS9; median fold change 0.05, IQR 0.02–0.43) from clonal to subclonal stages.

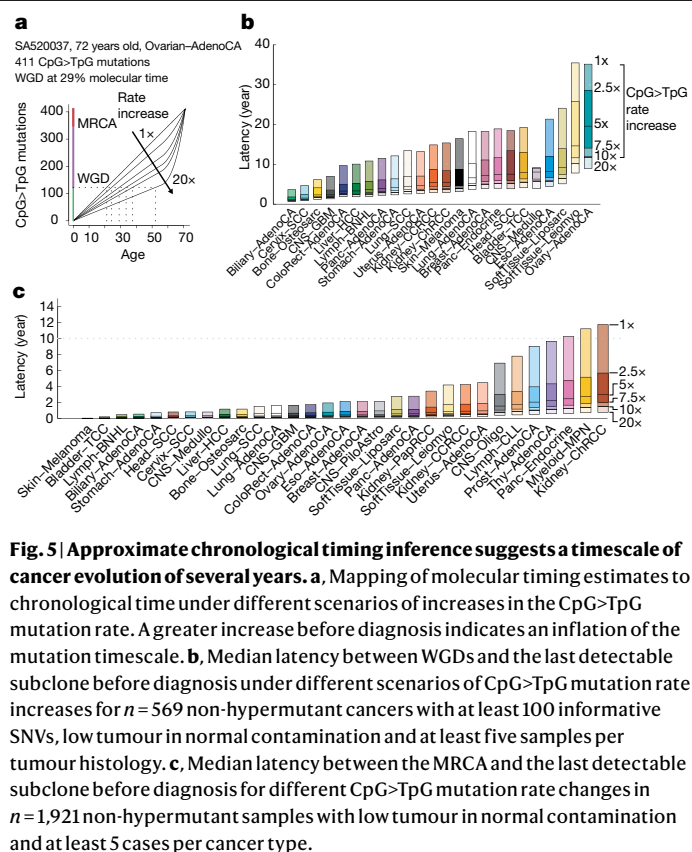
Some mutational processes tend to increase throughout cancer evolution. For example, we see that APOBEC mutagenesis (SBS2 and SBS13) increases in many cancer types from the early to late clonal stages (median fold change 2.0, IQR 0.8–3.6), as does a newly described signature SBS38 (median fold 3.6, IQR 1.8–11). Signatures of defective mismatch repair (SBS6, 14, 15, 20, 21, 26 and 44) increase from clonal to subclonal stages (median fold 1.8, IQR 1.2–3.0).

### Chronological time estimates

The molecular timing data presented above do not measure the occurrence of events in chronological time. If the rate at which mutations are acquired per year in each sample was constant, the chronological time would simply be the product of the estimated molecular timing and age at diagnosis. However, this relation will be nonlinear if the mutation rate changes over time, and is inflated by acquired mutational processes, as suggested by the analysis in the previous section. Some of these issues can be mitigated by counting only mutations contributed by endogenous and less variable mutational processes, such as CpG-to-TpG mutations (hereafter CpG>TpG) caused by spontaneous deamination of 5-methyl-cytosine to thymine at CpG dinucleotides, which have been proposed as a molecular clock<sup>12</sup>. Our supplementary analysis suggests that, although the baseline CpG>TpG mutation rate in cancers is very close to that in normal cells, there appears to be a moderate increase (1–10 times, adding between 20 and 40% of mutations) in cancers (Extended Data Fig. 8). As this shifts chronological timing estimates, we model different scenarios of the evolution of the CpG>TpG mutation rate (Fig. 5a).

Applying this logic to time WGDs, which yield sufficient numbers of CpG>TpG mutations, demonstrates that they occur several years and possibly even a decade or more before diagnosis in some cancer types, under a range of scenarios of mutation rate increase (Fig. 5b, Extended Data Fig. 9). A notable example is ovarian adenocarcinoma, which appears to have a median latency of more than 10 years. This holds true even under a scenario of a CpG>TpG rate increase of 20-fold, which would be far beyond the 7.5-fold rate increase observed in matched primary and relapse samples<sup>39</sup> (Extended Data Fig. 8f). Notably, these results suggest WGD may occur throughout the entire female reproductive life (Extended Data Fig. 9b). The latency between the MRCA and the last detectable subclone is shorter, typically several months to years (Fig. 5c).

These timescales of cancer evolution are further supported by the fact that progression of most known precancerous lesions to carcinomas usually spans many years, if not decades<sup>40–45</sup>. Our data corroborate these timescales and extend them to cancer types without detectable premalignant conditions, raising the hope that these tumours could also be detected in less malignant stages.



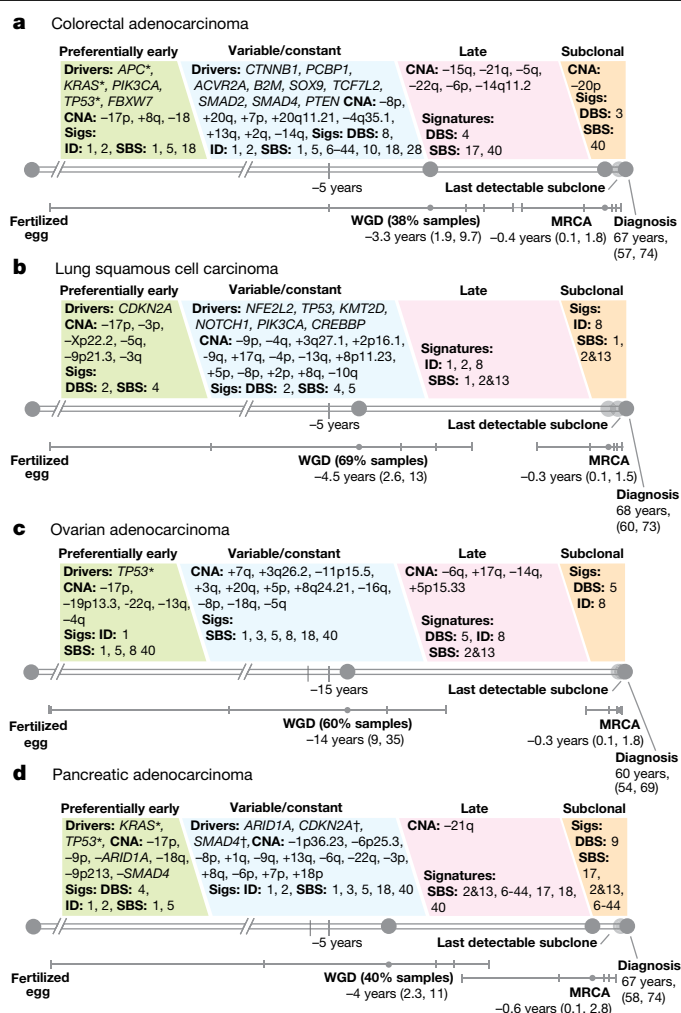
**Fig. 5 | Approximate chronological timing inference suggests a timescale of cancer evolution of several years.** **a**, Mapping of molecular timing estimates to chronological time under different scenarios of increases in the CpG>TpG mutation rate. A greater increase before diagnosis indicates an inflation of the mutation timescale. **b**, Median latency between WGDs and the last detectable subclone before diagnosis under different scenarios of CpG>TpG mutation rate increases for  $n = 569$  non-hypermutant cancers with at least 100 informative SNVs, low tumour in normal contamination and at least five samples per tumour histology. **c**, Median latency between the MRCA and the last detectable subclone before diagnosis for different CpG>TpG mutation rate changes in  $n = 1,921$  non-hypermutant samples with low tumour in normal contamination and at least 5 cases per cancer type.

## Discussion

To our knowledge, our study presents the first large-scale genome-wide reconstruction of the evolutionary history of cancers, reconstructing both early (pre-cancer) and later stages of 38 cancer types. This is facilitated by the timing of copy number gains relative to all other events in the genome, through multiplicity and clonal status of co-amplified point mutations. However, several limitations exist (Supplementary Information). Perhaps most importantly, molecular timing is based on point mutations and is therefore subject to changes in mutation rate. Notably, healthy tissues acquire point mutations at rates not too dissimilar from those seen in cancers, particularly when considering only endogenous mutational processes, and furthermore, some tissues are riddled with microscopic clonal expansions of driver gene mutations<sup>5-9,11</sup>. This is direct evidence that the life history of almost every cell in the human body, including those that develop into cancer, is driven by somatic evolution.

Together, the data presented here enable us to draw approximate timelines summarizing the typical evolutionary history of each cancer type (Fig. 6, Supplementary Information for all other cancer types). These make use of the qualitative timing of point mutations and copy number alterations, as well as signature activities, which can be interleaved with the chronological estimates of WGD and the appearance of the MRCA.

It is remarkable that the evolution of practically all cancers displays some level of order, which agrees very well with, and adds much detail to, established models of cancer progression<sup>35,46</sup>. For example, *TP53* with accompanying 17p deletion is one of the most frequent initiating mutations in a variety of cancers, including ovarian cancer, in which it is the hallmark of its precancerous precursor lesions<sup>47</sup>. Furthermore, the list of typically early drivers includes most other highly recurrent cancer genes, such as *KRAS*, *TERT* and *CDKN2A*, indicating a preferred role in early and possibly even pre-cancer evolution. This initially constrained set of genes broadens at later stages of cancer development,



**Fig. 6 | Typical timelines of tumour development.** **a-d**, Timelines representing the length of time, in years, between the fertilized egg and the median age of diagnosis for colorectal adenocarcinoma (**a**), squamous cell lung cancer (**b**), ovarian adenocarcinoma (**c**) and pancreatic adenocarcinoma (**d**). Real-time estimates for major events, such as WGD and the emergence of the MRCA, are used to define early, variable, late and subclonal stages of tumour evolution approximately in chronological time. The range of chronological time estimates according to varying clock mutation acceleration rates is shown as well, with tick marks corresponding to 1x, 2.5x, 5x, 7.5x, 10x and 20x. Driver mutations and copy number alterations (CNA) are shown in each stage according to their preferential timing, as defined by relative ordering. Mutational signatures (Sigs) that, on average, change over the course of tumour evolution, or are substantially active but not changing, are shown in the epoch in which their activity is greatest. DBS, double base substitution; SBS, single base substitutions. Where applicable, lesions with a known timing from the literature are annotated; dagger symbols denotes events that were found to have a different timing; asterisk symbol denotes events that agree with our timing.

suggesting an epistatic fitness landscape canalizing the first steps of cancer evolution. Over time, as tumours evolve, they follow increasingly diverse paths driven by individually rare driver mutations, and by copy number alternations. However, none of these trends is absolute, and the evolutionary paths of individual tumours are highly variable, showing that cancer evolution follows trends, but is far from deterministic.

Our study sheds light on the typical timescales of in vivo tumour development, with initial driver events seemingly occurring up to decades before diagnosis, demonstrating how cancer genomes are shaped by a lifelong process of somatic evolution, with fluid boundaries between normal ageing processes<sup>5-11</sup> and cancer evolution.

Nevertheless, the presence of genetic aberrations with such long latency raises hopes that aberrant clones could be detected early, before reaching their full malignant potential.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1907-7>.

- Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).
- Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
- Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
- Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Preprint at bioRxiv <https://doi.org/10.1101/505685> (2018).
- Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
- Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
- Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
- Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
- Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
- Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
- Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
- Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
- Durinck, S. et al. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* **1**, 137–143 (2011).
- Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol.* **19**, 95 (2018).
- Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell* **173**, 611–623 (2018).
- Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
- Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
- Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
- Brastianos, P. K. et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* **5**, 1164–1177 (2015).
- Papaemmanuil, E. et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–3627 (2013).
- Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
- Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1965-x> (2020).
- Alexandrov, L. B. The repertoire of mutational signatures in human cancer. *Nature* <https://doi.org/10.1038/s41586-020-1943-3> (2020).
- Keogh, M. J. et al. High prevalence of focal and multi-focal somatic genetic variants in the human brain. *Nat. Commun.* **9**, 4257 (2018).
- Heim, S. et al. Trisomy 7 and sex chromosome loss in human brain tissue. *Cytogenet. Cell Genet.* **52**, 136–138 (1989).
- Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278–282 (2009).
- Sheltzer, J. M. et al. Single-chromosome gains commonly function as tumor suppressors. *Cancer Cell* **31**, 240–255 (2017).
- Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
- Cross, W. et al. The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol.* **2**, 1661–1672 (2018).
- Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
- Gibson, W. J. et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat. Genet.* **48**, 848–855 (2016).

- Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184 (2017).
- Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
- Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
- Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
- Patch, A.-M. et al. Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).
- Bostwick, D. G. & Qian, J. High-grade prostatic intraepithelial neoplasia. *Mod. Pathol.* **17**, 360–379 (2004).
- Brenner, H. et al. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. *Gut* **56**, 1585–1589 (2007).
- Gazdar, A. F. & Brambilla, E. Preneoplasia of lung cancer. *Cancer Biomark.* **9**, 385–396 (2010).
- Sanders, M. E., Schuyler, P. A., Dupont, W. D. & Page, D. L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* **103**, 2481–2484 (2005).
- Schlecht, N. F. et al. Human papillomavirus infection and time to progression and regression of cervical intraepithelial neoplasia. *J. Natl. Cancer Inst.* **95**, 1336–1343 (2003).
- Whitson, M. J. & Falk, G. W. Predictors of progression to high-grade dysplasia or adenocarcinoma in Barrett's esophagus. *Gastroenterol. Clin. North Am.* **44**, 299–315 (2015).
- Bardeesy, N. & DePinho, R. A. Pancreatic cancer biology and genetics. *Nat. Rev. Cancer* **2**, 897–909 (2002).
- Folkens, A. K. et al. A candidate precursor to pelvic serous cancer (p53 signature) and its prevalence in ovaries and fallopian tubes from women with BRCA mutations. *Gynecol. Oncol.* **109**, 168–173 (2008).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

## PCAWG Evolution & Heterogeneity Working Group

Stefan C. Drento<sup>3,4,6</sup>, Ignaty Leshchiner<sup>5</sup>, Moritz Gerstung<sup>1,2,3</sup>, Clemency Jolly<sup>4</sup>, Kerstin Haase<sup>4</sup>, Maxime Tarabichi<sup>3,4</sup>, Jeff Wintersinger<sup>8,9</sup>, Amit G. Deshwar<sup>8,9</sup>, Kaixian Yu<sup>11</sup>, Santiago Gonzalez<sup>7</sup>, Yulia Rubanova<sup>8,9</sup>, Geoff Macintyre<sup>16</sup>, David J. Adams<sup>3</sup>, Pavana Anur<sup>10</sup>, Rameen Beroukhi<sup>5,31</sup>, Paul C. Boutros<sup>8,25,26</sup>, David D. Bowtell<sup>27</sup>, Peter J. Campbell<sup>3</sup>, Shaolong Cao<sup>11</sup>, Elizabeth L. Christie<sup>19,27</sup>, Marek Cmero<sup>19,20</sup>, Yupeng Cun<sup>34</sup>, Kevin J. Dawson<sup>3</sup>, Jonas Demeulemeester<sup>4,21</sup>, Nilgun Donmez<sup>17,18</sup>, Ruben M. Drews<sup>16</sup>, Roland Eils<sup>12,13</sup>, Yu Fan<sup>11</sup>, Matthew Fittall<sup>4</sup>, Dale W. Garsed<sup>19,27</sup>, Gad Getz<sup>5,28,29,30</sup>, Gavin Ha<sup>5</sup>, Marcin Imielinski<sup>22,23</sup>, Lara Jerman<sup>114</sup>, Yuan Ji<sup>15,33</sup>, Kortine Kleinheinz<sup>12,13</sup>, Juhee Lee<sup>24</sup>, Henry Lee-Six<sup>3</sup>, Dimitri G. Livitz<sup>5</sup>, Salem Malikic<sup>17,18</sup>, Florian Markowitz<sup>16</sup>, Inigo Martincorena<sup>3</sup>, Thomas J. Mitchell<sup>3,7</sup>, Ville Mustonen<sup>35</sup>, Layla Oesper<sup>42</sup>, Martin Peifer<sup>34</sup>, Myron Peto<sup>10</sup>, Benjamin J. Raphael<sup>43</sup>, Daniel Rosebrock<sup>5</sup>, S. Cenik Sahinalp<sup>18,32</sup>, Adriana Salcedo<sup>25</sup>, Matthias Schlesner<sup>12</sup>, Steven Schumacher<sup>5</sup>, Subhajit Sengupta<sup>15</sup>, Ruian Shi<sup>8</sup>, Seung Jun Shin<sup>11,44</sup>, Oliver Spiro<sup>5</sup>, Lincoln D. Stein<sup>25</sup>, Ignacio Vázquez-García<sup>37</sup>, Shankar Vembu<sup>8</sup>, David A. Wheeler<sup>45</sup>, Tsun-Po Yang<sup>34</sup>, Xiaotong Yao<sup>22,23</sup>, Ke Yuan<sup>16,36</sup>, Hongtu Zhu<sup>11</sup>, Wenyi Wang<sup>11</sup>, Quaid D. Morris<sup>8,9</sup>, Paul T. Spellman<sup>10</sup>, David C. Wedge<sup>6,38</sup> & Peter Van Loo<sup>4,21</sup>

<sup>42</sup>Department of Computer Science, Carleton College, Northfield, MN, USA. <sup>43</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA. <sup>44</sup>Korea University, Seoul, South Korea. <sup>45</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

## Methods

### Dataset

The PCAWG series consists of 2,778 tumour samples (2,703 white listed, 75 grey listed) from 2,658 donors. All samples in this dataset underwent whole-genome sequencing (minimum average coverage 30× in the tumour, 25× in the matched normal samples), and were processed with a set of project-specific pipelines for alignment, variant calling, and quality control<sup>4</sup>. Copy number calls were established by combining the output of six individual callers into a consensus using a multi-tier approach, resulting in a copy number profile, a purity and ploidy value and whether the tumour has undergone a WGD (Supplementary Information). Consensus subclonal architectures have been obtained by integrating the output of 11 subclonal reconstruction callers, after which all SNVs, indels and structural variants are assigned to a mutation cluster using the MutationTimer.R approach (Supplementary Information). Driver calls have been defined by the PCAWG Driver Working Group<sup>4</sup>, and mutational signatures are defined by the PCAWG Signatures Working Group<sup>24</sup>. A more detailed description can be found in Supplementary Information, section 1.

Data accrual was based on sequencing experiments performed by individual member groups of the ICGC and TCGA, as described in an associated study<sup>4</sup>. As this is a meta-analysis of existing data, power calculations were not performed and the investigators were not blinded to cancer diagnoses.

### Timing of gains

We used three related approaches to calculate the timing of copy number gains (see Supplementary Information, section 2). In brief, the common feature is that the expected VAF of a mutation ( $E$ ) is related to the underlying number of alleles carrying a mutation according to the formula:  $E[X] = nmfp / [N(1 - \rho) + Cp]$ , in which  $X$  is the number of reads,  $n$  denotes the coverage of the locus, the mutation copy number  $m$  is the number of alleles carrying the mutation (which is usually inferred),  $f$  is the frequency of the clone carrying the given mutation ( $f = 1$  for clonal mutations),  $N$  is the normal copy number (2 on autosomes, 1 or 2 for chromosome X and 0 or 1 for chromosome Y),  $C$  is the total copy number of the tumour, and  $\rho$  is the purity of the sample.

The number of mutations  $n_m$  at each allelic copy number  $m$  then informs about the time when the gain has occurred. The basic formulae for timing each gain are, depending on the copy number configuration:

$$\text{Copy number } 2 + 1: T = 3n_2 / (2n_2 + n_1)$$

$$\text{Copy number } 2 + 2: T = 2n_2 / (2n_2 + n_1)$$

$$\text{Copy number } 2 + 0: T = 2n_2 / (2n_2 + n_1)$$

in which 2 + 1 refers to major and minor copy number of 2 and 1, respectively. Methods differ slightly in how the number of mutations present on each allele are calculated and how uncertainty is handled (Supplementary Information).

### Timing of mutations

The mutation copy number  $m$  and the clonal frequency  $f$  is calculated according to the principles indicated above. Details can be found in Supplementary Information, section 2. Mutations with  $f = 1$  are denoted as 'clonal', and mutations with  $f < 1$  as 'subclonal'. Mutations with  $f = 1$  and  $m > 1$  are denoted as 'early clonal' (co-amplified). In cases with  $f = 1$ ,  $m = 1$  and  $C > 2$ , mutations were annotated as 'late clonal', if the minor copy number was 0, otherwise 'clonal' (unspecified).

### Timing of driver mutations

A catalogue of driver point mutations (SNVs and indels) was provided by the PCAWG Drivers and Functional Interpretation Group<sup>4</sup>. The timing

category was calculated as above. From the four timing categories, the odds ratios of early/late clonal and clonal (early, late or unspecified clonal)/subclonal were calculated for driver mutations against the distribution of all other mutations present in fragments with the same copy number composition in the samples with each particular driver. The background distribution of these odds ratios was assessed with 1,000 bootstraps (Supplementary Information, section 4.1).

### Integrative timing

For each pair of driver point mutations and recurrent copy number alterations, an ordering was established (earlier, later or unspecified). The information underlying this decision was derived from the timing of each driver point mutation, as well as from the timing status of clonal and subclonal copy number segments. These tables were aggregated across all samples and a sports statistics model was employed to calculate the overall ranking of driver mutations. A full description is given in Supplementary Information, section 4.2.

### Timing of mutational signatures

Mutational trinucleotide substitution signatures, as defined by the PCAWG Mutational Signatures Working Group<sup>24</sup>, were fit to samples with observed signature activity, after splitting point mutations into either of the four epochs. A likelihood ratio test based on the multinomial distribution was used to test for differences in the mutation spectra between time points. Time-resolved exposures were calculated using non-negative linear least squares. Full details are given in Supplementary Information, section 5.

### Real-time estimation of WGD and MRCA

CpG>TpG mutations were counted in an NpCpG context, except for skin-melanoma, in which CpCpG and TpCpG were excluded owing to the overlapping UV mutation spectrum. For visual comparison, the number of mutations was scaled to the effective genome size, defined as the  $1/\text{mean}(m_i/C_i)$ , in which  $m_i$  is the estimated number of allelic copies of each mutation, and  $C_i$  is the total copy number at that locus, thereby scaling to the final copy number and the time of change.

A hierarchical Bayesian linear regression was fit to relate the age at diagnosis to the scaled number of mutations, ensuring positive slope and intercept through a shared gamma distribution across cancer types.

For tumours with several time points, the set of mutations shared between diagnosis and relapse ( $n_D$ ) and those specific to the relapse ( $n_R$ ) was calculated. The rate acceleration was calculated as:  $a = n_R/n_D \times t_D/t_R$ . This analysis was performed separately for all substitutions and for CpG>TpG mutations.

On the basis of these analyses, a typical increase of 5× for most cancer types was chosen, with a lower value of 2.5× for brain cancers and a value of 7.5× for ovarian cancer.

The correction for transforming an estimate of a copy number gain in mutation time into chronological time depends not only on the rate acceleration, but also on the time at which this acceleration occurred. As this is generally unknown, we performed Monte Carlo simulations of rate accelerations spanning an interval of 15 years before diagnosis, corresponding roughly to 25% of time for a diagnosis at 60 years of age, noting that a 5× rate increase over this duration yields an offset of about 33% of mutations, compatible with our data. Subclonal mutations were assumed to occur at full acceleration. The proportion of subclonal mutations was divided by the number of identified subclones, thus conservatively assuming branching evolution. Full details are given in Supplementary Information, section 6.

### Cancer timelines

The results from each of the different timing analyses are combined in timelines of cancer evolution for each tumour type (Fig. 6 and Supplementary Information). Each timeline begins at the fertilized egg, and spans up to the median age of diagnosis within each cohort. Real-time



# Article

estimates for WGD and the MRCA act as anchor points, allowing us to roughly map the four broadly defined time periods (early clonal, intermediate, late clonal and subclonal) to chronological time during a patient's lifespan. Specific driver mutations or copy number alterations can be placed within each of these time frames based on their ordering from the league model analysis. Signatures are shown if they typically change over time (95% confidence intervals of mean change not overlapping 0), and if they are strongly active (contributing at least 10% mutations to one time point). Signatures are shown on the timeline in the epoch of their greatest activity. Where an event found in our study has a known timing in the literature, the agreement is annotated on the timeline; with an asterisk denoting an agreed timing, and dagger symbol denoting a timing that is different to our results. Full details are given in Supplementary Information, section 7.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA PCAWG Consortium are described elsewhere<sup>4</sup> and available for download at <https://dcc.icgc.org/releases/PCAWG>. Further information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access information that could potentially identify participants, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic SNVs derived from TCGA donors, researchers will also need to obtain dbGaP authorization. Datasets used and results presented in this study, including timing estimates for copy number gains, chronological estimates of WGD and MRCA, as well as mutation signature changes, are described in Supplementary Note 3 and are available at <https://dcc.icgc.org/releases/PCAWG/evolution-heterogeneity>.

## Code availability

The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at <https://dockstore.org/search?search=pcawg> under the GNU General Public License v3.0, which allows for reuse and distribution. Analysis code presented in this study is available through the GitHub repository <https://github.com/PCAWG-11/Evolution>. This archive contains

relevant software and analysis workflows as submodules, which include code for timing copy number gains, point mutations and mutation signatures, real-time timing and evolutionary league model analysis, as well as scripts to generate the figures presented: CancerTiming (v.3.1.8), MutationTimeR (v.0.1), PhylogicNDT (v.1.1) and a series of custom scripts (v.1.0), with detailed versions of other packages used.

**Acknowledgements** We thank H. Lee-Six and L. Moore for sharing data on mutation burden in normal tissues. This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202) and the Wellcome Trust (FC001202). This project was enabled through the Crick Scientific Computing STP and through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (grant number MR/L016311/1). M.T. and J.D. are postdoctoral fellows supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie grant agreement number 747852-SIOMICs and 703594-DECODE). J.D. is a postdoctoral fellow of the FWO. F.M., G.M. and K. Yuan acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. G.M., K. Yuan and F.M. were funded by CRUK core grants C14303/A17197 and A19274. S. Sengupta and Y.J. are supported by NIH R01 CA132897. S.M. is supported by the Vanier Canada Graduate Scholarship. S.C.S. is supported by the NSERC Discovery Frontiers Project, "The Cancer Genome Collaboratory" and NIH Grant GM108308. H.Z. is supported by grant NIMH086633 and an endowed Bao-Shan Jing Professorship in Diagnostic Imaging. W.W. is supported by the US National Cancer Institute (1R01 CA183793 and P30 CA016672). P.T.S. was supported by U24CA210957 and U24CA143799. D.C.W. is funded by the Li Ka Shing foundation. P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

**Author contributions** M.G., C.J., I.L., S.G., P.A., D.R., D.G.L., P.T.S. and P.V.L. performed timing of point mutations and copy number gains. S.G. and M.G. performed qualitative timing of driver point mutations and analyses of synchronous gains, L.J. timed secondary copy number gains. I.L., T.J.M., D.R., D.G.L., D.C.W. and G.G. performed relative timing of somatic driver events and implemented integrative models. C.J., Y.R., M.G., Q.D.M. and P.V.L. performed timing of mutational signatures. M.G. performed real-time estimation of whole-genome duplication and subclonal diversification. S.G. assessed mutation rates in relapsed samples. C.J., M.G., I.L., Y.R., D.R. and P.V.L. constructed cancer timelines. M.G., C.J., I.L., S.C.D., S.G., T.J.M., Y.R., P.A., J.D., P.C.B., D.D.B., V.M., Q.D.M., P.T.S., D.C.W. and P.V.L. interpreted the results. S.C.D., I.L., J.W., A.D., I.V.-G., K. Yuan, G.M., M.P., S.M., N.D., K. Yu, S. Sengupta, K.H., M.T., J.D., D.G.L., D.R., J.L., M.C., S.C.S., Y.J., F.M., V.M., H.Z., W.W., Q.D.M., D.C.W. and P.V.L. performed subclonal architecture analysis. S.C.D., I.L., K.K., V.M., M.P., X.Y., D.G.L., S. Schumacher, R.B., M.I., M.S., D.C.W. and P.V.L. performed copy number analysis. J.W., S.C.D., I.L., K.H., D.G.L., K.K., D.R., D.C.W., Q.D.M. and P.V.L. derived a consensus of copy number analysis results. K. Yu, M.T., A.D., S.C.D., I.L., D.C.W., M.G., P.V.L., Q.D.M. and W.W. derived a consensus of subclonal architecture results. Y.F. and W.W. contributed to subclonal mutation calls. P.T.S., D.C.W. and P.V.L. coordinated the study. M.G., C.J., P.T.S., Y.R., I.L., Q.D.M., D.C.W. and P.V.L. wrote the manuscript, which all authors approved. S.C.D., I.L., M.G., C.J., K.H., M.T., J.W., A.G.D., K. Yu, S.G., Y.R. and G.M. in the PCAWG Evolution & Heterogeneity Working Group contributed equally. W.W., Q.D.M., P.T.S., D.C.W. and P.V.L. in the PCAWG Evolution & Heterogeneity Working Group jointly supervised the work.

**Competing interests** R.B. owns equity in Amprezza Therapeutics. G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect and POLYSOLVER. I.L. is a consultant for PACT Pharma. B.J.R. is a consultant at and has ownership interest (including stock and patents) in Medley Genomics. All other authors declare no competing interests.

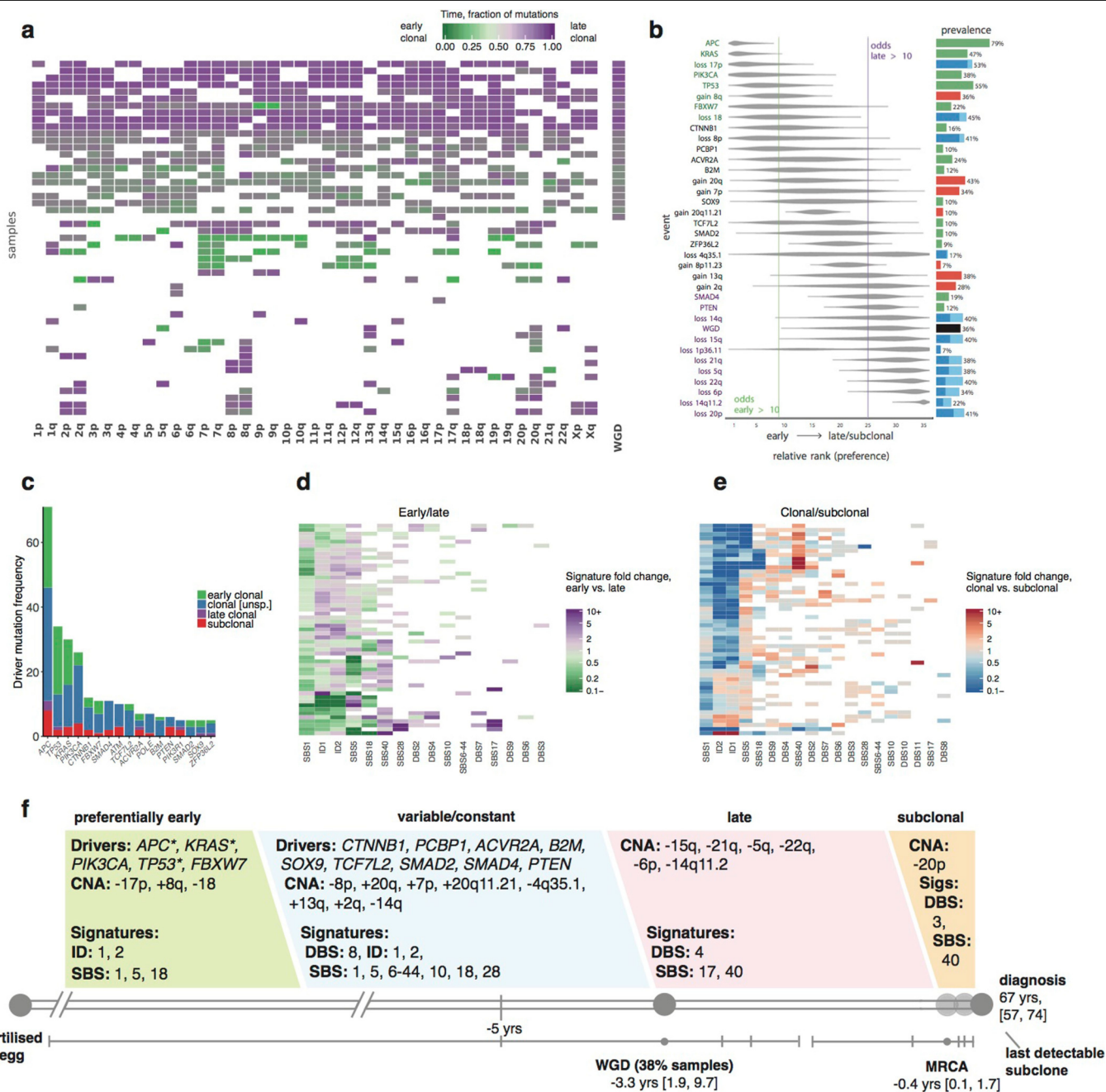
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1907-7>.

**Correspondence and requests for materials** should be addressed to M.G. or P.V.L.

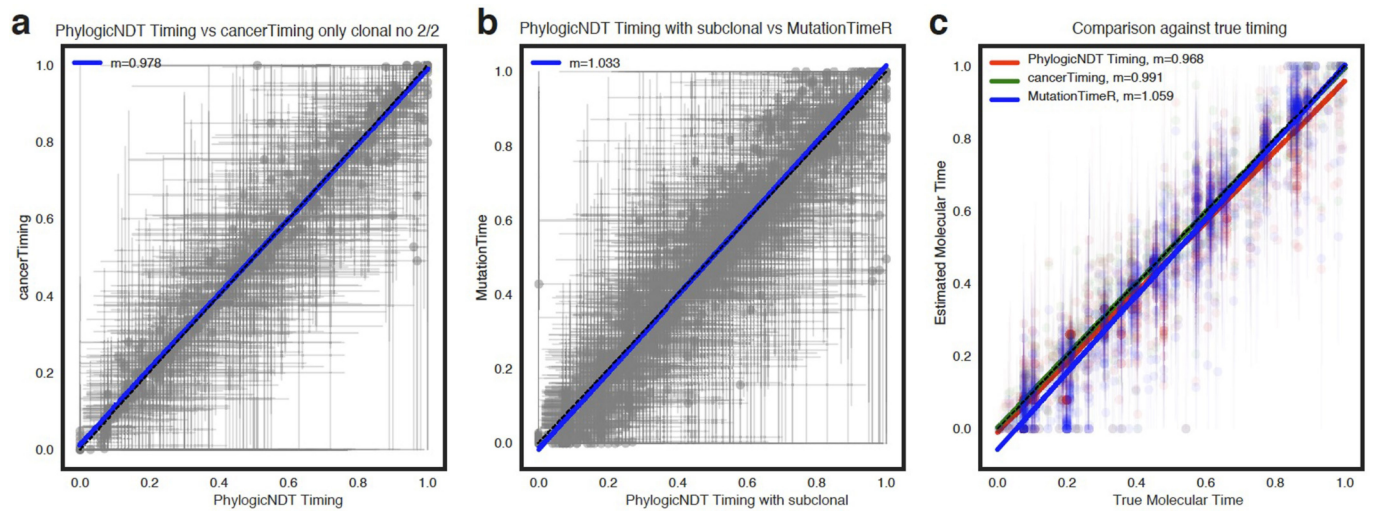
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



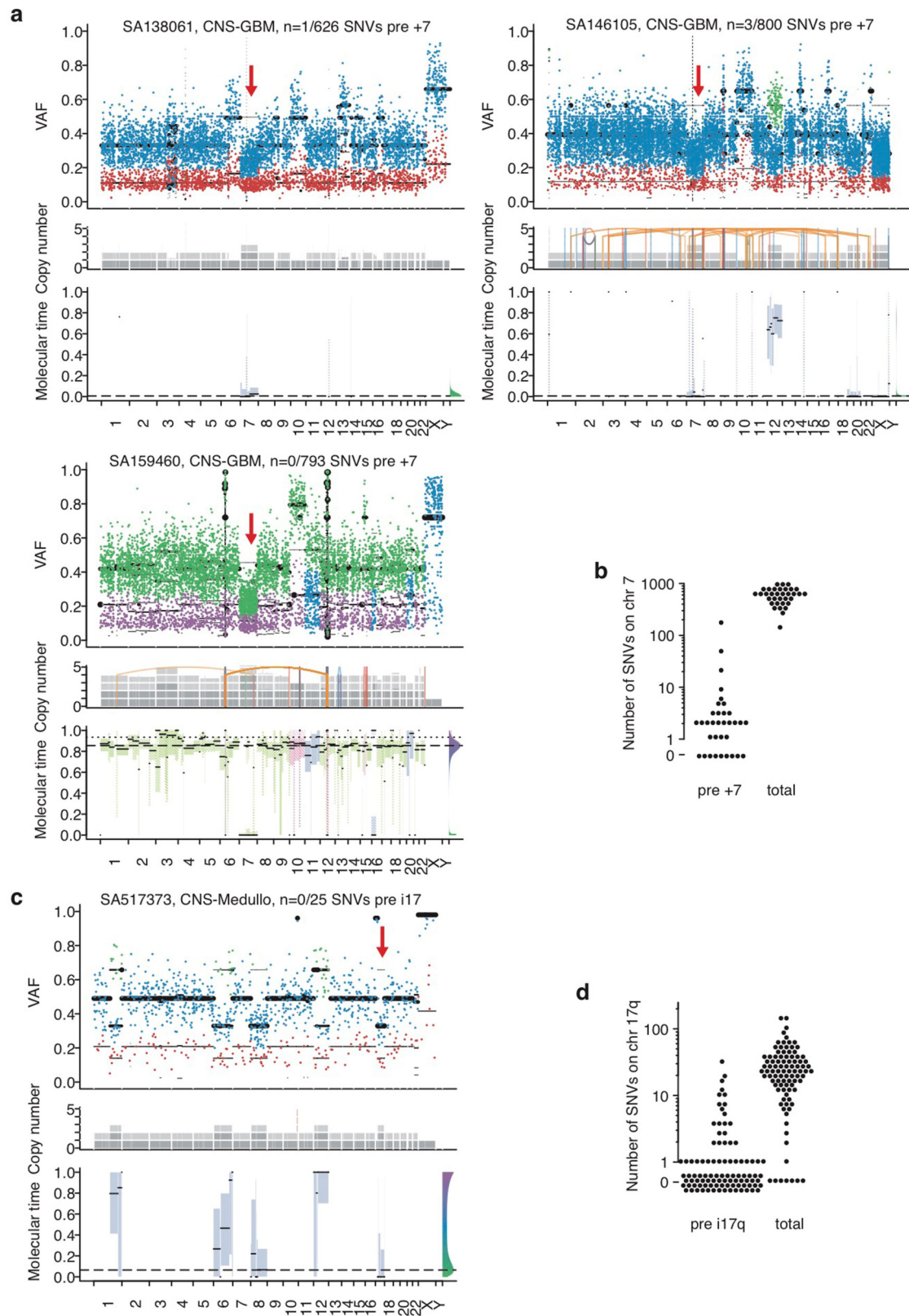


**Extended Data Fig. 1 | Summary of all results obtained for colorectal adenocarcinoma ( $n = 60$ ) as an example. a**, Clustered heat maps of mutational timing estimates for gained segments, per patient. Colours as indicated in main text: green represents early clonal events, purple represents late clonal. **b**, Relative ordering of copy number events and driver mutations across all samples. **c**, Distribution of mutations across early clonal, late clonal and subclonal stages, for the most common driver genes. A maximum of 10 driver genes are shown. **d**, Clustered mutational signature fold changes between early

clonal and late clonal stages, per patient. Green and purple indicate, respectively, a signature decrease and increase in late clonal from early clonal mutations. Inactive signatures are coloured white. **e**, As in **d** but for clonal versus subclonal stages. Blue indicates a signature decrease and red an increase in subclonal from clonal mutations. **f**, Typical timeline of tumour development. Similar result summaries for all other cancer types can be found in the Supplementary Information (pages 46–77).

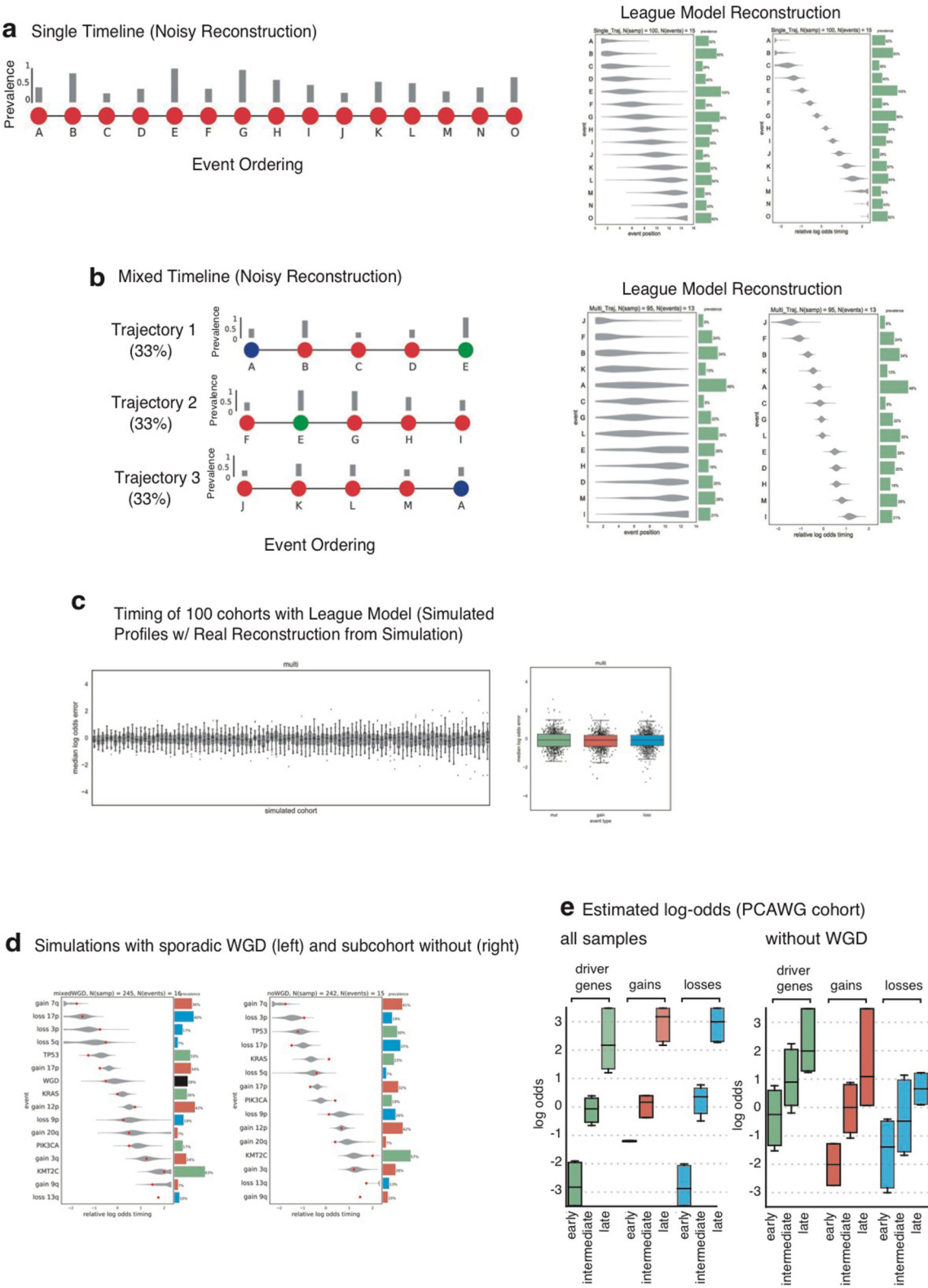


**Extended Data Fig. 2 | Comparison of methods used for timing of individual copy number gains. a, b,** Pairwise comparison of the three approaches for timing individual copy number gains. **c,** Comparison using simulated data, showing high concordance.



**Extended Data Fig. 3 | Early copy number gains in brain cancers. a,** Three illustrative examples of glioblastoma with trisomy 7. The red arrow depicts the expected VAF cluster of point mutations preceding trisomy 7, which usually contains less than three SNVs. **b,** Distributions of the number of SNVs preceding trisomy 7 and total number of mutations on chromosome (chr) 7 in

$n = 34$  GBM samples with trisomy 7. **c,** Medulloblastoma example with isochromosome 17q. **d,** Distributions of SNVs on 17q in  $n = 95$  samples with isochromosome 17q; 74 out of 95 samples have less than 1 SNV preceding the isochromosome.

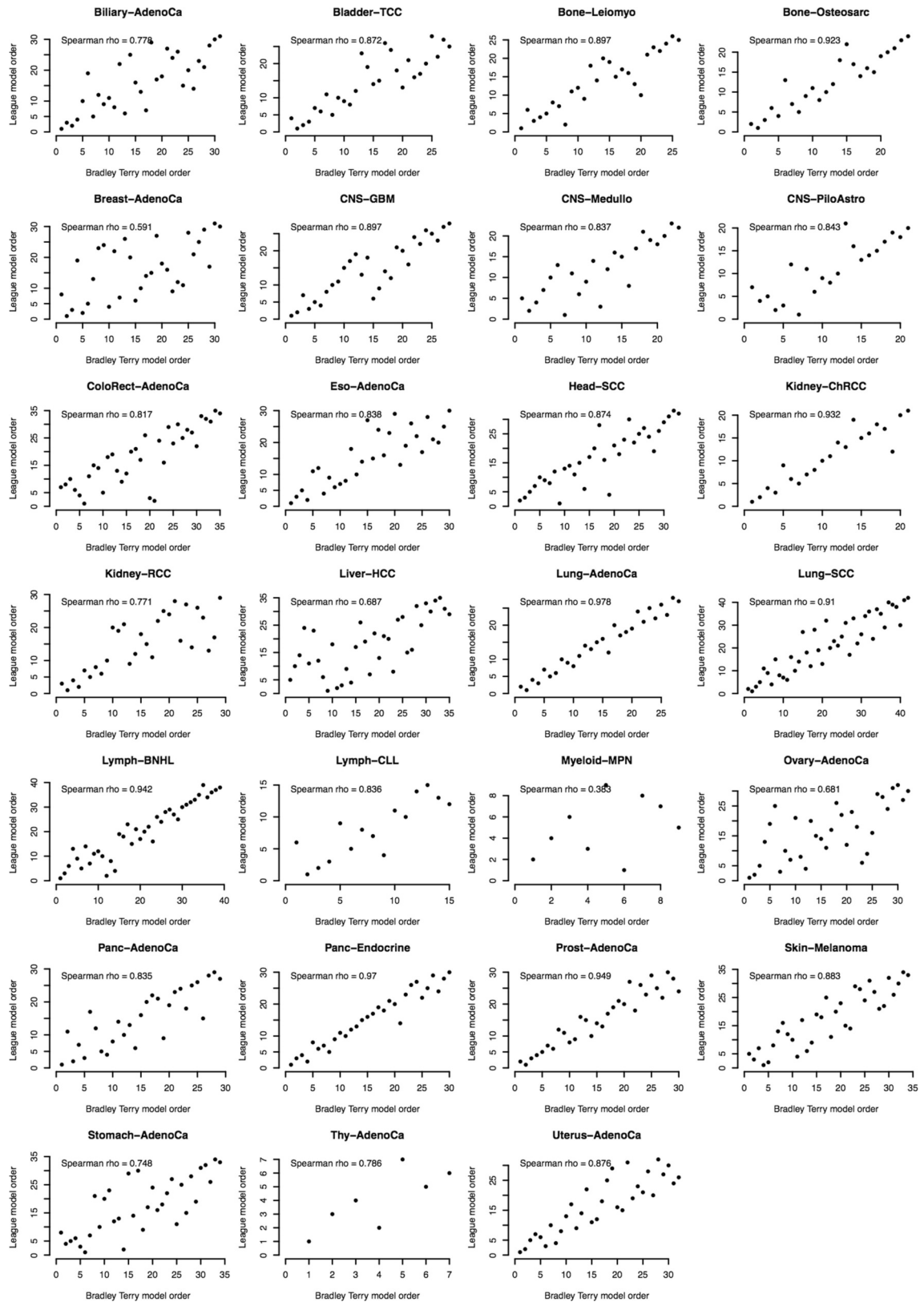


Extended Data Fig. 4 | See next page for caption.

**Extended Data Fig. 4 | Validation of relative ordering model reconstruction based on simulated cohorts of whole-genome samples.** **a**, Relative ordering model (PhylogicNDT LeagueModel) results for a simulated cohort of samples ( $n = 100$ ) from a single generalized relative order of events (with varied prevalence) showing high concordance with the true trajectory. Probability distributions show the uncertainty of timing for specific events in the cohort. **b**, Relative ordering model results on a simulated cohort of samples ( $n = 95$ ) from a complex mixture of trajectories with different order of events showing high concordance with the expected average trajectory. **c**, Estimation of accuracy of the relative ordering model reconstruction by simulation of a set of 100 cohorts ( $n(\text{samples}) = 100$ ) with random trajectory mixtures and quantifying the distance in log odds early/late from perfect ordering. For the vast majority of events (even with low number of occurrences in the cohort),

the log odds error does not exceed 1, confirming that very few events would switch between timing categories. The inset box corresponds to the first and third quartiles of the distribution, the horizontal line indicates the median and whiskers include data within  $1.5 \times$  the IQR from the box. **d**, Simulated data show concordant timing in cohorts with WGD ( $n = 245$ ). Exclusion of samples with WGD (right,  $n = 242$ ) introduces only a mild drop in accuracy, indicating that WGD is beneficial but not necessary for the reconstruction. Red dot = true rank. **e**, Estimated log odds in observed data including WGD (left,  $n = 245$ ) and without (right,  $n = 242$ ), across different mutation types. The inset box corresponds to the first and third quartiles of the distribution, the horizontal line indicates the median and whiskers include data within  $1.5 \times$  the IQR from the box.

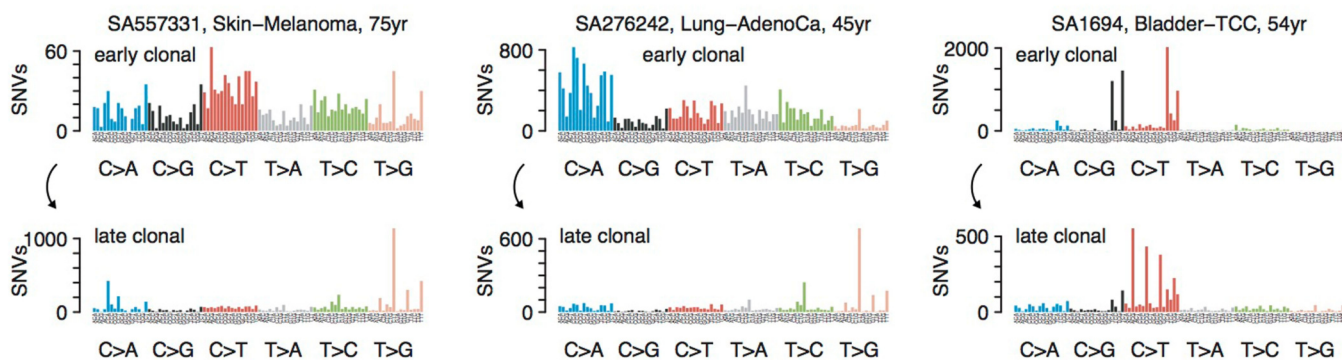




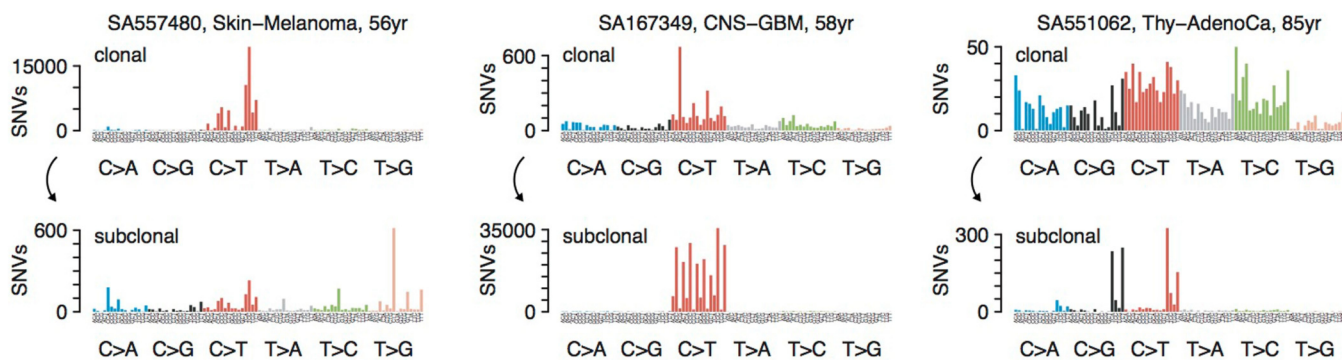
**Extended Data Fig. 5 | Correlation between the league model and Bradley-Terry model ordering.** Direct comparison for each tumour type of the league and Bradley-Terry models for determining the order of recurrent somatic

mutations and copy number events. Axes indicate the ordered events observed in the respective tumour types. Correlation is quantified by Spearman's rank correlation coefficient. A total of  $n = 756$  ordered events are shown.

**a** Examples: early to late clonal mutation spectrum changes

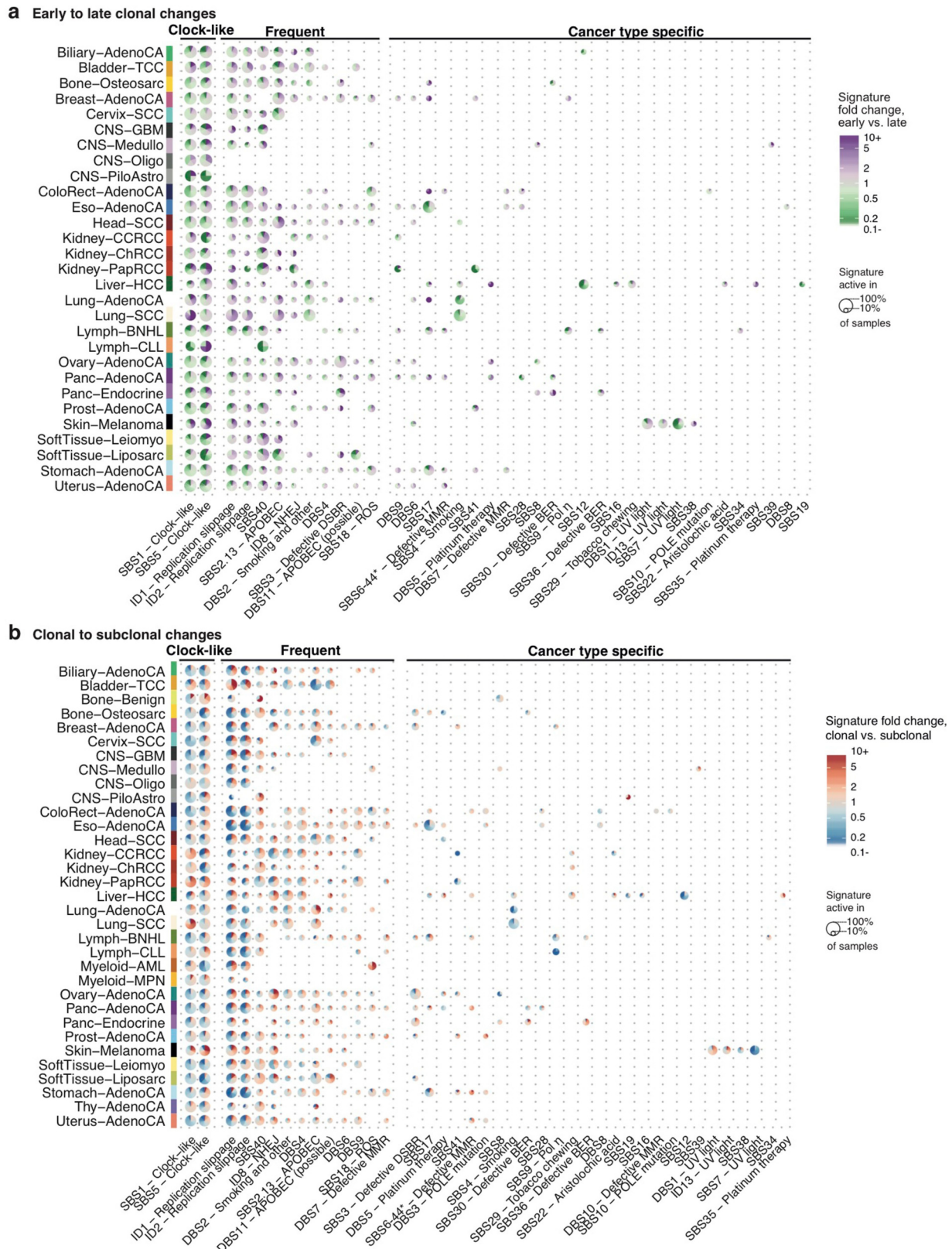


**b** Examples: clonal to subclonal mutation spectrum changes



**Extended Data Fig. 6 | Examples of mutation spectrum changes across tumour evolution. a.** Three examples of tumours with substantial changes between mutation spectra of early (top) and late (bottom) clonal time points.

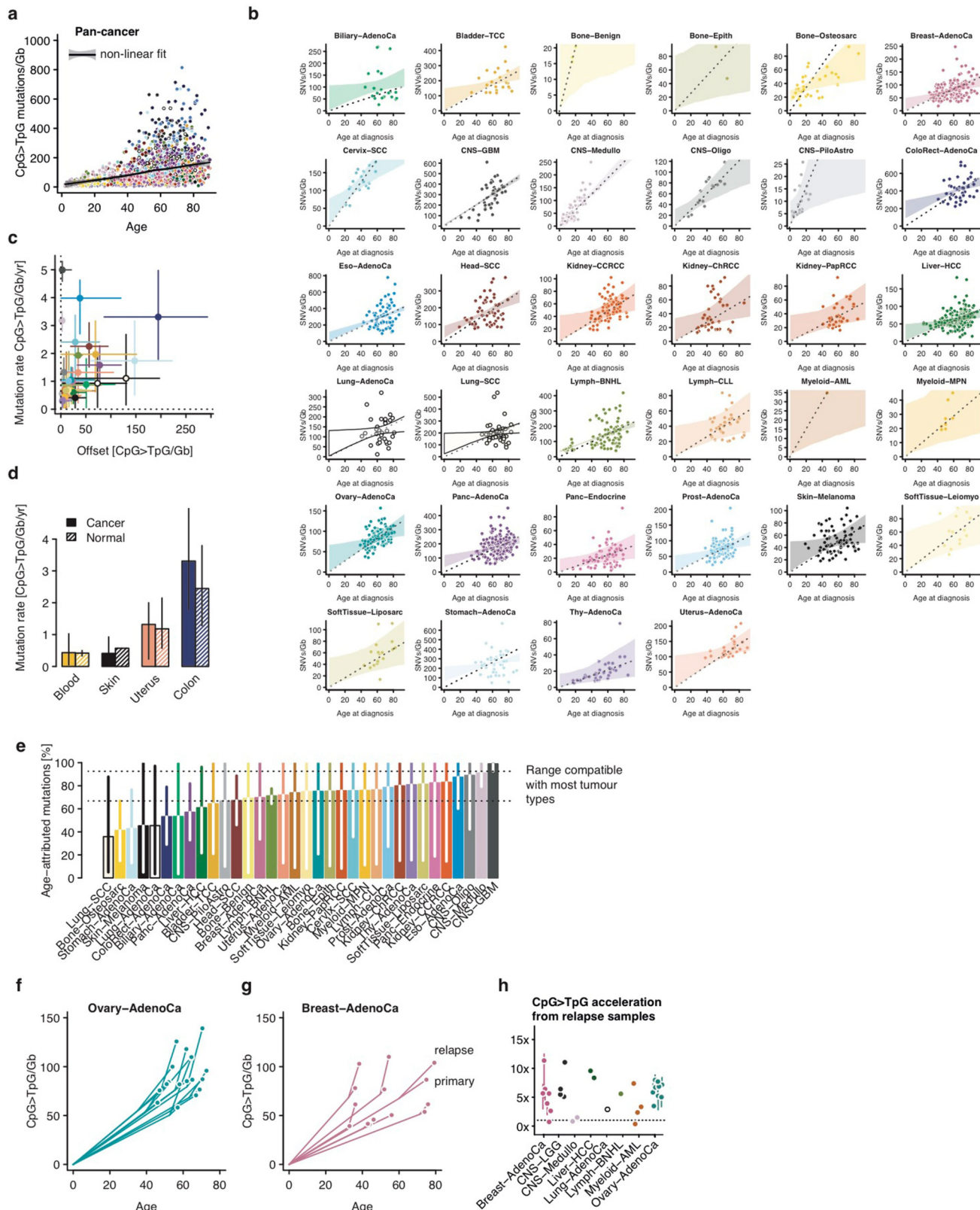
**b.** Three examples of tumours with substantial changes between mutation spectra of clonal (top) and subclonal (bottom) time points.



**Extended Data Fig. 7 | Overview of early-to-late clonal and clonal-to-subclonal signature changes across tumour types. a, b, Pie charts representing signature changes per cancer type for early-to-late clonal signature changes (a) and clonal-to-subclonal signature changes (b). Signatures that decrease between early and late are coloured green; signatures that increase are purple. The size of each pie chart represents the frequency of**

each signature. Signatures are split into three categories: (1) clock-like, comprising the putative clock signatures 1 and 5; (2) frequent, which are signatures present in ten or more cancer types; and (3) cancer-type specific, which are in fewer than ten cancer types and are often limited to specific cohorts.

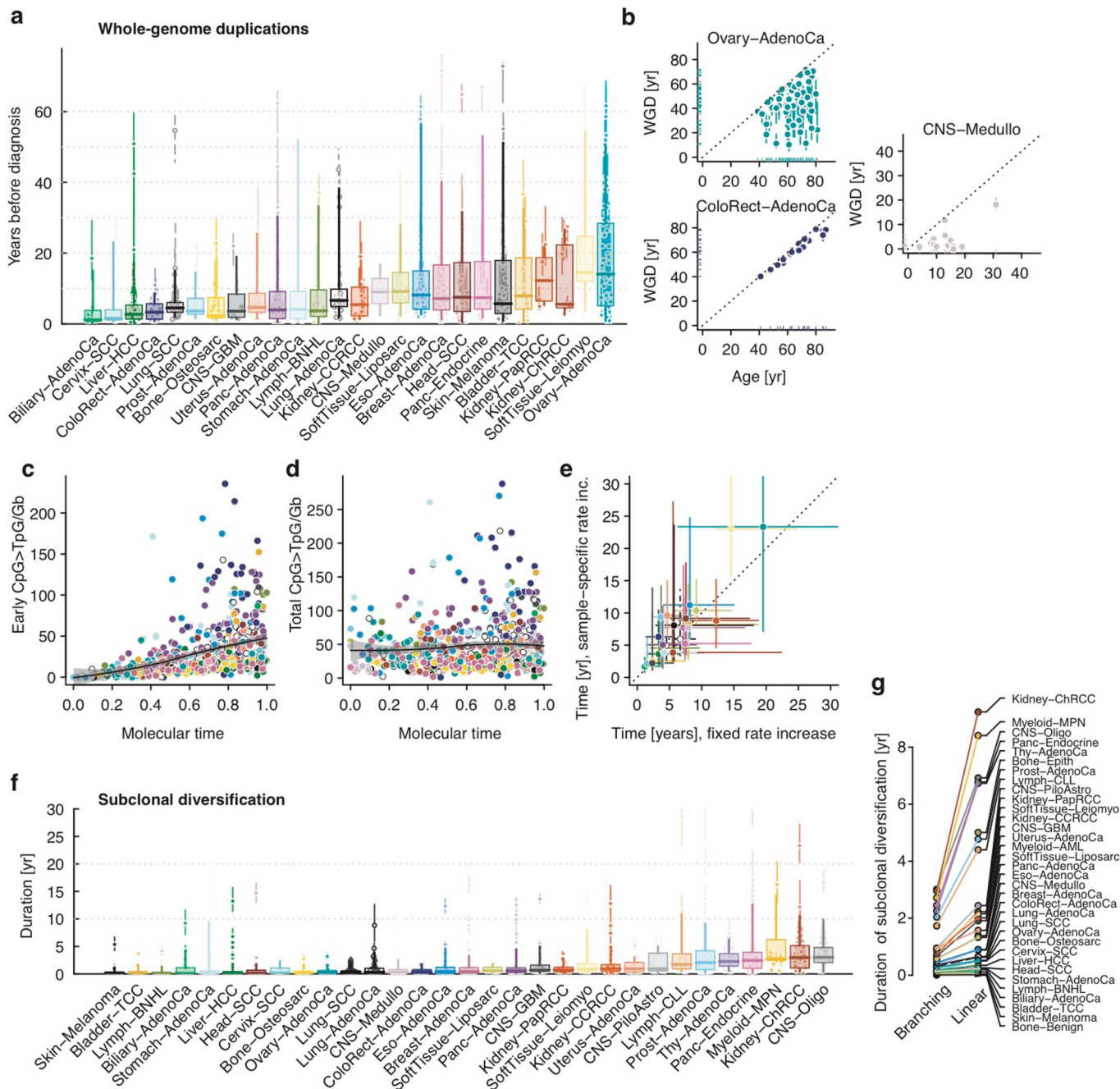




**Extended Data Fig. 8** | See next page for caption.

**Extended Data Fig. 8 | Age-dependent mutation burden and relapse samples indicate near-normal CpG>TpG mutation rate in cancer, with moderate acceleration during carcinogenesis.** **a**, Across all cancer samples, a predominantly linear accumulation of CpG>TpG mutations (scaled to copy number) is observed over time, as measured by the age at diagnosis. **b**, Cancer-specific analysis of the CpG>TpG mutation burden as a function of age at diagnosis for  $n=1,978$  samples of 34 informative cancer types. The dotted line denotes the median mutations per year (that is, not offset), and shading denotes the 95% credible interval of a hierarchical Bayesian linear regression model across all data points. Slope and intercepts are drawn for each cancer type from a gamma distribution, respectively; inference was done by Hamiltonian Monte Carlo sampling. **c**, Maximum a posteriori estimates of rate and offset for 34 cancer types with 95% credible intervals as defined in **b**. **d**, Mutation rate inferred from cancer as in **b** and from selected normal tissue sequencing studies of  $n=140$  normal haematopoietic stem cells,  $n=1$  normal skin sample,  $n=182$  samples from normal endometrium, and  $n=445$  normal colonic crypts; error bars denote the 95% confidence interval. **e**, Median fraction of mutations attributed to linear age-dependent accumulation, based on estimates from **b** and the age at diagnosis for each sample. Error bars denote the 95% credible interval. **f, g**, CpG>TpG mutations per gigabase for ovarian cancer (**f**) and breast cancer (**g**) samples with matched primary and relapse samples. **h**, Increase in CpG>TpG mutation rate inferred from paired primary and relapse samples for six cancer types. Bars denote the range of the rate increase for different scenarios of copy number evolution, assuming ploidy changes have occurred prior (upper value) or posterior (lower value) to the branching between primary and relapse sample.





**Extended Data Fig. 9 | Real-time estimates indicate long latencies for some samples caused by the absence of early mutations.** **a**, Time of WGD for  $n = 571$  individual patients, split by tumour type with an estimated mutation rate increase of 5%, except for ovary-adenocarcinoma (7.5%) and CNS (2.5%). Error bars represent 80% confidence intervals, reflecting uncertainty stemming from the number of mutations per segment and onset of the rate increase. Box plots demarcate the quartiles and median of the distribution with whiskers indicating 5% and 95% quantiles. **b**, Scatter plots showing the time of diagnosis (x axis) and inferred time of WGD (y axis) with error bars as in **a**. **c**, Scatter plot of early (co-amplified) CpG>TpG mutations (y axis) as a function of the mutational time estimate of WGD (x axis). The black line denotes a nonlinear loess fit with 95% confidence interval. Colours define the cancer type as in **a**. **d**, Total

CpG>TpG mutations (y axis) as a function of the mutation time estimate of WGD (x axis). Colours and fit as in **c**. Early molecular timing is thus caused by a depletion of early CpG>TpG mutations, rather than an inflation of late CpG>TpG mutations. **e**, Estimated median WGD latency of  $n = 571$  patients as in **a** for fixed (x axis) versus patient specific rate increases, depending on the observed CpG>TpG mutation burden, allowing for a higher (up to 10 $\times$ ) mutation rate increase in samples with more mutations (y axis). Error bars denote the IQR. **f**, Timing of subclonal diversification using CpG>TpG mutations in  $n = 1,953$  individual patients. Box plots and error bars for data points as in **a**. **g**, Comparison of the median duration of subclonal diversification per cancer type assuming branching and linear phylogenies.

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

## Software and code

Policy information about [availability of computer code](#)

Data collection	Data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at <a href="https://www.overture.bio">https://www.overture.bio</a> . Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to <a href="mailto:contact@overture.bio">contact@overture.bio</a> .
Data analysis	The PCAWG workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: <a href="https://dockstore.org/search?labels.value.keyword=pcawg&amp;searchMode=files">https://dockstore.org/search?labels.value.keyword=pcawg&amp;searchMode=files</a> . Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACESeq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBbA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15. Analysis code presented in this study is available through the github repository <a href="https://github.com/PCAWG-11/Evolution">https://github.com/PCAWG-11/Evolution</a> . This archive contains relevant software and analysis workflows as submodules, including code for timing copy number gains, point mutations and mutation signatures, real-time timing, and evolutionary league model analysis, as well as scripts to generate the figures presented.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

WGS somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization. All results presented in this study, including timing estimates for copy number gains, real time estimates of WGD and MRCA, as well as mutation signature activities, are available at <https://www.synapse.org/#!Synapse:syn14193595>.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014.
Data exclusions	After quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine). Hypermethylated and samples with normal contamination were excluded for chronological inferences in this study, as described in the Supplementary Methods.
Replication	In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement. The accuracy of inferences in this study was assessed using simulations and by applying three different algorithms for the timing of copy number gains (Extended Data Figure 2), as well as two different algorithms for the temporal ordering of driver mutations (Extended Data Figure 5).
Randomization	N/A - This exploratory study did not contain a randomization step
Blinding	N/A - This exploratory study did not contain a blinded analysis

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

### Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<p>Patient-by-patient clinical data are provided in Extended Data Table 1 of the marker paper for the PCAWG consortium. Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.</p>
Recruitment	<p>Patients were recruited by the participating centres following local protocols.</p>
Ethics oversight	<p>The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.</p>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Genomic basis for RNA alterations in cancer

<https://doi.org/10.1038/s41586-020-1970-0>

Received: 29 March 2018

Accepted: 11 December 2019

Published online: 5 February 2020

Open access

PCAWG Transcriptome Core Group<sup>1,35</sup>, Claudia Calabrese<sup>2,35</sup>, Natalie R. Davidson<sup>3,4,5,6,7,35</sup>, Deniz Demircioğlu<sup>8,9,35</sup>, Nuno A. Fonseca<sup>2,35</sup>, Yao He<sup>10,35</sup>, André Kahles<sup>3,4,6,7,35</sup>, Kjong-Van Lehmann<sup>3,4,6,7,35</sup>, Fenglin Liu<sup>10,35</sup>, Yuichi Shiraishi<sup>11,35</sup>, Cameron M. Soulette<sup>12,35</sup>, Lara Urban<sup>2,35</sup>, Liliana Greger<sup>2</sup>, Siliang Li<sup>13,14</sup>, Dongbing Liu<sup>13,14</sup>, Marc D. Perry<sup>15,16</sup>, Qian Xiang<sup>15</sup>, Fan Zhang<sup>10</sup>, Junjun Zhang<sup>15</sup>, Peter Bailey<sup>17</sup>, Serap Erkek<sup>18</sup>, Katherine A. Hoadley<sup>19</sup>, Yong Hou<sup>13,14</sup>, Matthew R. Huska<sup>20</sup>, Helena Kilpinen<sup>21</sup>, Jan O. Korbel<sup>18</sup>, Maximilian G. Marin<sup>12</sup>, Julia Markowski<sup>20</sup>, Tannistha Nandi<sup>9</sup>, Qiang Pan-Hammarström<sup>13,22</sup>, Chandra Sekhar Pedomallu<sup>23,28,29</sup>, Reiner Siebert<sup>24</sup>, Stefan G. Stark<sup>3,4,6,7</sup>, Hong Su<sup>13,14</sup>, Patrick Tan<sup>9,25</sup>, Sebastian M. Waszak<sup>18</sup>, Christina Yung<sup>15</sup>, Shida Zhu<sup>13,14</sup>, Philip Awadalla<sup>15,26</sup>, Chad J. Creighton<sup>27</sup>, Matthew Meyerson<sup>23,28,29</sup>, B. F. Francis Ouellette<sup>30</sup>, Kui Wu<sup>13,14</sup>, Huanming Yang<sup>13</sup>, PCAWG Transcriptome Working Group<sup>1</sup>, Alvis Brazma<sup>2,36\*</sup>, Angela N. Brooks<sup>12,23,28,36\*</sup>, Jonathan Göke<sup>9,31,36</sup>, Gunnar Rätsch<sup>3,4,5,6,7,36\*</sup>, Roland F. Schwarz<sup>2,20,32,33,36</sup>, Oliver Stegle<sup>2,18,33,36</sup>, Zemin Zhang<sup>10,36</sup> & PCAWG Consortium<sup>34</sup>

Transcript alterations often result from somatic changes in cancer genomes<sup>1</sup>. Various forms of RNA alterations have been described in cancer, including overexpression<sup>2</sup>, altered splicing<sup>3</sup> and gene fusions<sup>4</sup>; however, it is difficult to attribute these to underlying genomic changes owing to heterogeneity among patients and tumour types, and the relatively small cohorts of patients for whom samples have been analysed by both transcriptome and whole-genome sequencing. Here we present, to our knowledge, the most comprehensive catalogue of cancer-associated gene alterations to date, obtained by characterizing tumour transcriptomes from 1,188 donors of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)<sup>5</sup>. Using matched whole-genome sequencing data, we associated several categories of RNA alterations with germline and somatic DNA alterations, and identified probable genetic mechanisms. Somatic copy-number alterations were the major drivers of variations in total gene and allele-specific expression. We identified 649 associations of somatic single-nucleotide variants with gene expression in *cis*, of which 68.4% involved associations with flanking non-coding regions of the gene. We found 1,900 splicing alterations associated with somatic mutations, including the formation of exons within introns in proximity to Alu elements. In addition, 82% of gene fusions were associated with structural variants, including 75 of a new class, termed ‘bridged’ fusions, in which a third genomic location bridges two genes. We observed transcriptomic alteration signatures that differ between cancer types and have associations with variations in DNA mutational signatures. This compendium of RNA alterations in the genomic context provides a rich resource for identifying genes and mechanisms that are functionally implicated in cancer.

For a more extensive study of cancer genome alterations, particularly in non-coding regions, the PCAWG project was formed to analyse the large number of whole-genome samples that were contributed to the ICGC and TCGA projects<sup>5</sup>. Individual projects did not use the same methods for key analyses; therefore, a major focus for each of the 16 PCAWG Working Groups was the unified analysis of the PCAWG data. For example, the PCAWG Technical Working Group led raw data collection, realignment of whole-genome sequencing data and implemented core somatic mutation calling pipelines<sup>5</sup>. Other PCAWG working groups focused on unified analyses of copy-number variation<sup>6</sup>, structural variants<sup>7,8</sup>, germline variants<sup>5</sup>,

mutational signatures<sup>9</sup> and identification of driver genes<sup>8</sup>, among others<sup>5</sup>. Here, we report the joint analysis of available matched transcriptome and genome profiling for 1,188 samples from 27 tumour types by the PCAWG Transcriptome Working Group<sup>5</sup>, providing the largest, to our knowledge, resource of RNA phenotypes and their underlying genetic changes in cancer so far (Extended Data Fig. 1, Methods, Supplementary Results, Supplementary Table 23). We demonstrate the importance of transcriptomics data in understanding how different dimensions of specific DNA alterations contribute to carcinogenesis and map out the landscape of cancer-related RNA alterations.

A list of affiliations appears at the end of the paper.



## Cancer-specific germline *cis*-eQTLs

To investigate the underlying mechanisms of different types of RNA alteration, we first focused on changes in the gene expression level (Extended Data Fig. 2). We initially considered common germline variants (minor allele frequency  $\geq 1\%$ ) proximal to individual genes ( $\pm 100$  kb), and mapped expression quantitative trait loci (eQTL) across the cohort (Extended Data Fig. 3, Supplementary Table 1). This pan-cancer analysis identified 3,532 genes with an eQTL (false discovery rate (FDR)  $\leq 5\%$ , hereafter denoted eGenes) (Supplementary Table 2), enriched in proximal regions of transcription start sites (TSSs) (Extended Data Fig. 3).

To identify cancer-specific regulatory variants, we compared our eQTLs to eQTLs from the Genotype-Tissue Expression (GTEx) project<sup>10</sup>, adopting previous strategies to assess eQTL replication<sup>11</sup>, and probed lead eQTL variants for marginal significance in GTEx tissues ( $P \leq 0.01$ , Bonferroni-adjusted). Although most lead variants could be detected in GTEx samples (3,110 out of 3,532 eQTL variants), we identified 422 eQTLs that did not correspond to GTEx tissues, which suggests cancer-specific regulation (Extended Data Fig. 4, Supplementary Table 3). The corresponding eQTL lead variants were enriched for heterochromatic regions (Fig. 1a). Overall, this analysis revealed that the germline framework of gene expression regulation is largely conserved in cancer tissues.

## Somatic *cis*-eQTLs in non-coding regions

Previous studies have described the landscape of non-coding mutations in cancer<sup>1</sup>, particularly in promoter regions, and also their regulatory effects on gene expression<sup>12,13</sup>. Here, we looked at possible somatic DNA changes, across the whole genome, that underlie alterations in gene expression. We estimated local mutation burdens by aggregating single-nucleotide variants (SNVs) in 2-kb intervals adjacent to genes (flanking), as well as in exons and introns (Extended Data Figs. 2, 5, 6). Next, we decomposed the expression variation of individual genes, considering common mutation burdens in *cis*, as well as *cis* germline variants and somatic copy-number alterations (SCNAs). This identified SCNAs as the major driver of expression variation (17%), followed by somatic SNVs in gene flanking regions (1.8%) and germline variants (1.3%) (Fig. 1b).

We also tested for associations between all common mutation burdens and gene expression across the whole genome. We identified 649 genes with a somatic eQTL (FDR  $\leq 5\%$ ) (Supplementary Table 5). Of these, 11 associations were located in introns or exons of the respective eGene, including genes with known roles in the pathogenesis of specific cancers such as *CDK12* in ovarian cancer<sup>14</sup> and *IRF4* in chronic lymphocytic leukaemia<sup>15</sup> (Extended Data Figs. 7, 8). Most eQTLs (68.4%) involved associations with flanking non-coding mutation burdens (Extended Data Fig. 6e). Next, we considered eQTLs in flanking regions ( $n = 556$ ) and tested for enrichment in cell-type-specific annotations from the Epigenetics Roadmap<sup>16</sup>. This identified 13 enriched annotations (FDR  $\leq 10\%$ ) (Extended Data Fig. 9, Supplementary Table 6), including poised promoters, weak and active enhancers, and heterochromatin, but notably no enrichment for transcription-factor-binding sites (Supplementary Table 7). This enrichment in transcriptionally inactive regions may be due to an increased mutation rate in these regions (Extended Data Fig. 9), which has previously been reported in cancer<sup>17</sup>.

We also looked at the functional characterization of somatic eGenes and observed an enrichment for somatic eQTLs in bivalent promoters for cancer testis genes ( $P = 0.04$ , Fisher's exact test) such as *TEKT5*<sup>18</sup> (Fig. 1c, Extended Data Fig. 8h). Furthermore, we found a global enrichment (FDR  $\leq 10\%$ ) for Gene Ontology (GO) categories related to cell differentiation and developmental processes (Supplementary Table 8). Overall, somatic eQTL analysis identified mostly non-coding regions

associated with changes in local gene expression and, similar to cancer-specific germline eQTLs, showed enrichment for transcriptionally inactive regions such as heterochromatin.

## Expression and mutational signatures

Global variations in mutational patterns can be quantified using mutational signatures, which tag mutational processes specific to their tissue-of-origin and environmental exposures<sup>19</sup>. However, the extraction of mutational signatures is an intrinsically statistical process that requires a posteriori functional annotation. We performed a pan-cancer association analysis between genome-wide mutational signatures and gene expression levels to decipher the molecular processes that accompany the presence of mutational signatures.

We considered 28 mutational signatures derived using non-negative matrix factorization of context-specific mutation frequencies<sup>9</sup>. We tested for association between signature prevalence in donors and total gene expression, accounting for total mutational burden, cancer type, and other technical and biological confounders. This identified 1,176 genes associated with at least one signature (FDR  $\leq 10\%$ ) (Extended Data Fig. 10, Supplementary Table 19).

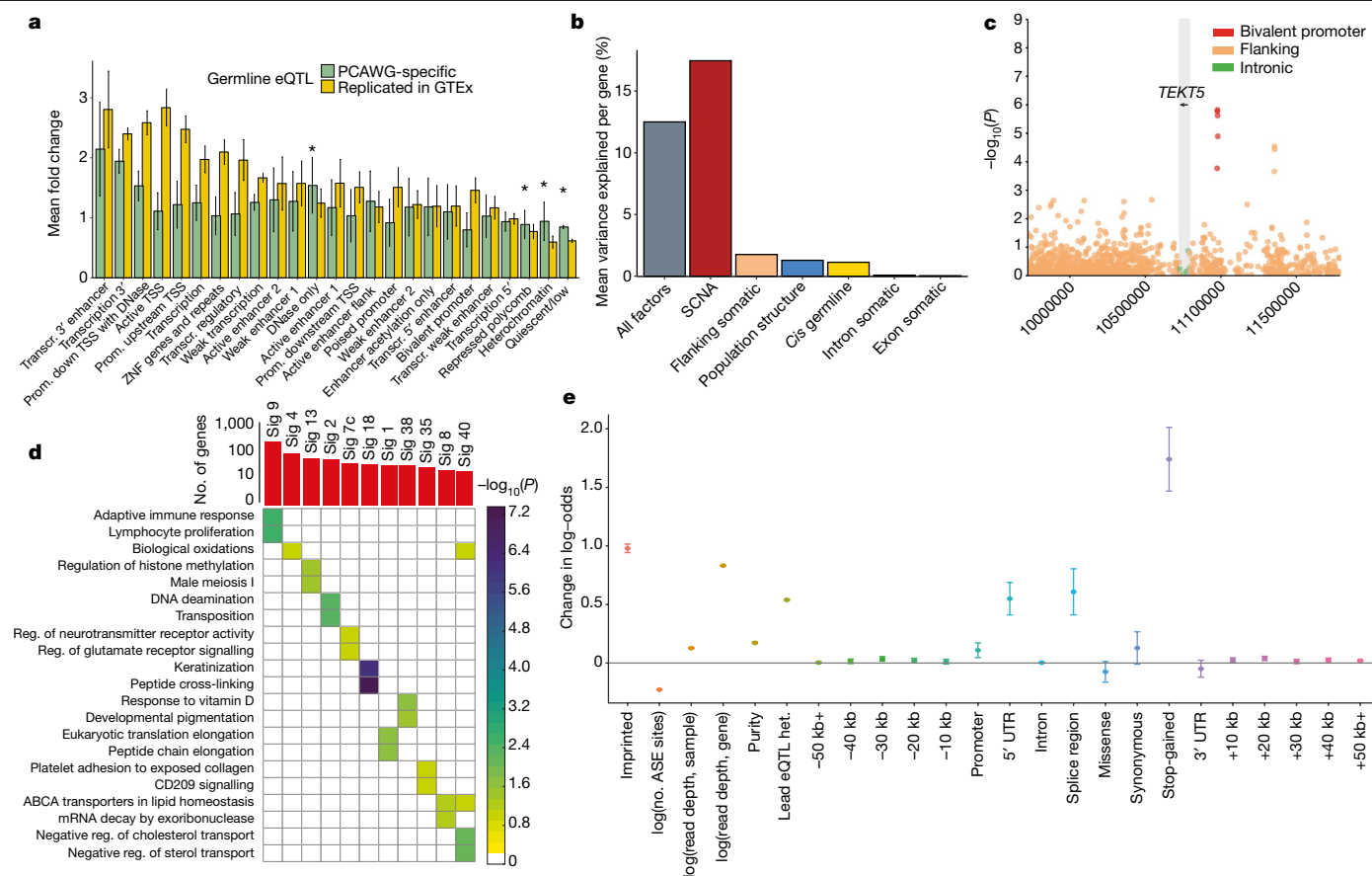
We considered 18 signatures with 20 or more associated genes for further annotation (Extended Data Fig. 11) and assessed enrichment using GO categories<sup>20</sup> and Reactome pathways<sup>21</sup>. We found that 11 signatures were enriched for at least one category (FDR  $\leq 10\%$ ) (Supplementary Table 19), revealing associations consistent with known and unknown aetiologies (Fig. 1d). For example, signature 38, which is correlated with the canonical UV signature 7 ( $r^2 = 0.375$ ,  $P = 5 \times 10^{-40}$ ) (Extended Data Fig. 11c), was linked to melanin processes (Fig. 1d). The synthesis of melanin causes oxidative stress to melanocytes<sup>22</sup>, and we found signature 38 associated with the oxidative-stress-promoting gene *TYR*<sup>23</sup> ( $P = 1.0 \times 10^{-4}$ ). A hallmark of signature 38 genes are C>A mutations, a typical product of reactive oxygen species<sup>24</sup>. This suggests that signature 38 may capture DNA damage that is indirectly caused by UV-induced oxidative damage after direct sun exposure<sup>25</sup>, with *TYR* as a possible mediator of the effect.

## Genomic basis of allelic expression

To analyse expression at the level of individual haplotypes, we tested for allelic expression imbalance (AEI) (FDR  $\leq 5\%$ , binomial test). We observed substantial differences in the fraction of genes with AEI between different types of cancer (Extended Data Fig. 12), and between cancer and the corresponding healthy tissues, with a high observed concordance between allelic imbalance at the DNA and RNA levels (Extended Data Fig. 13).

We used a logistic regression model to identify the determinants of AEI, accounting for known imprinting status<sup>26</sup>, the germline eQTL genotype, SCNAs and the weighted mutational burden of proximal somatic SNVs stratified into functional categories (Extended Data Fig. 2). In aggregate, SCNAs accounted for 84.3% of the total explained variation, which confirmed our findings from the somatic eQTL analysis, followed by germline eQTL lead variants (9.1%), somatic SNVs (4.9%) and imprinting status (1.7%) (Extended Data Fig. 14). Although cumulatively, non-coding variants were more relevant than coding variants, somatic protein-truncating variants ('stop-gained' variants) that triggered nonsense-mediated decay<sup>27</sup> were the most predictive individually. SNVs within splice regions, 5' untranslated regions (UTRs) and promoters were also strongly associated with the presence of AEI, and we observed a global trend of decreasing relevance of variants with increasing distance from the TSS (Fig. 1e, Extended Data Fig. 14).

Gene-centric attribution of AEI to individual sources of genetic variation (Supplementary Table 9) revealed an enrichment of somatically induced AEI in several known cancer-driver genes, as well as new candidates, such as the mismatch-repair-related gene *EXO1* that is associated



**Fig. 1 | Germline and somatic SNVs associated with expression.** **a**, Epigenetics Roadmap enrichment analysis, showing the average fold change in Roadmap factors across cell lines in PCAWG-specific eQTLs of the pan-analysis as well as eQTLs that replicate in GTEx tissues. \* $P < 0.05/25$ , one-sided Wilcoxon rank-sum test in PCAWG-specific eQTLs corrected for the number of Roadmap factors used (that is, 25). Data are mean and s.d. **b**, Variance component analysis for gene expression levels, showing the average proportion of variance explained by different germline and somatic factors for different sets of genes including the mean effect across all factors: (1) all genetic factors (germline and somatic); (2) SCNAs; (3) somatic variants in flanking regions; (4) population structure; (5) *cis*-germline effects; and (6) somatic intron and exon mutation effects. **c**, Manhattan plot showing nominal  $P$  values of association for *TEKT5* (highlighted in grey), considering flanking, intronic and exonic intervals.

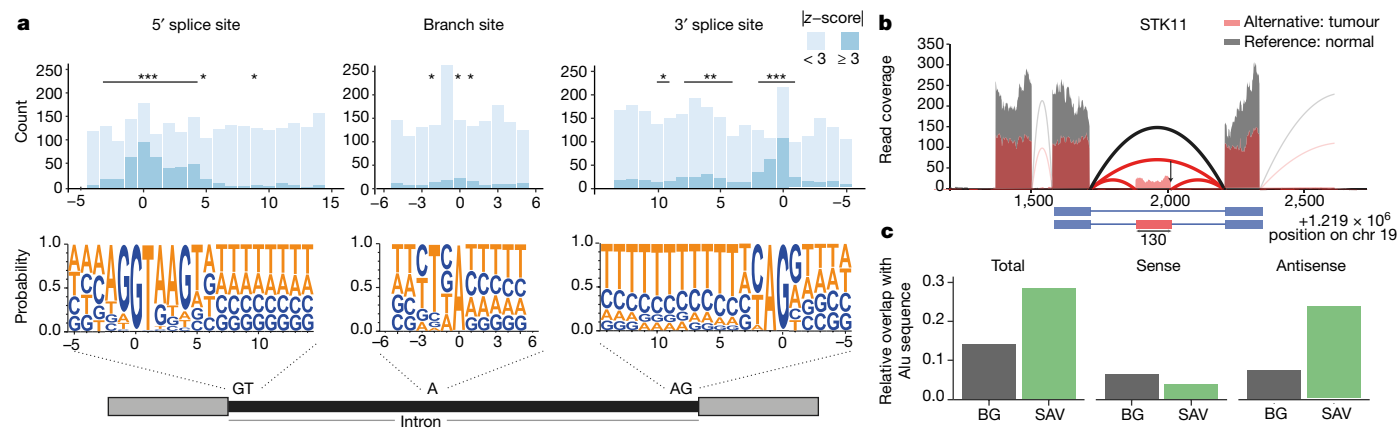
with survival in colorectal adenocarcinoma (log-rank  $P = 0.022$ , hazard ratio = 0.57) (Supplementary Results). We further observed a strong enrichment in the AEI score of cancer testis genes based on somatic SNVs only ( $\chi^2$  test  $P = 6 \times 10^{-3}$ ). In summary, we identify somatic and germline genetic variation that is associated with allele-specific dysregulation of genes across cancer types.

## Mutations associated with promoter usage

We considered promoter activity<sup>28–30</sup> as another molecular phenotype to study the effect of promoter mutations. Although cancer-specific alternative promoter usage has previously been shown<sup>28</sup>, the association of underlying genomic alterations with promoter activities have not been broadly explored. To estimate the activity of individual gene promoters, we combined the expression of isoforms initiated in TSSs that are identical or nearby, assuming that these are transcribed from the same promoter (Extended Data Fig. 15a–c). We divided promoters into three categories: (1) inactive promoters (activity  $< 1$  fragment per kilobase of transcript per million mapped reads (FPKM)), (2) major promoters (most active per gene) and (3) minor (all remaining) promoters,

The leading somatic burden is associated with increased *TEKT5* expression ( $P = 1.61 \times 10^{-6}$ ) and overlaps an upstream bivalent promoter (red dots; annotated in 81 Roadmap cell lines, including 8 embryonic stem cells, 9 embryonic-stem-cell-derived and 5 induced pluripotent stem-cell lines). **d**, Summary of significant associations between mutational signatures (Sig) and gene expression. Top, the total number of associated genes per signature (FDR  $\leq 10\%$ ). Bottom, enriched GO categories or Reactome pathways for genes associated with each signature (FDR  $\leq 10\%$ , significance level encoded in color,  $-\log_{10}$ -transformed adjusted  $P$  value). **e**, Standardized effect sizes on the presence of AEI, taking only SCNAs, germline eQTLs, coding and non-coding mutations into account. Data are the estimate and standard error of the estimate of the effect size.

and examined the rates of mutation across varying activity levels. We observed an increase in the number of mutations near the TSS of major promoters compared with minor or inactive promoters (Extended Data Fig. 15d). This pattern is most prominent in skin melanoma, in which it has been attributed to impaired nucleotide-excision repair (Extended Data Fig. 15e, f, k, l). The cancer type that shows the strongest deviation from this pattern is colorectal adenocarcinoma, which highlights the tissue-specificity of mutational patterns at promoters (Extended Data Fig. 15e, f, m, n). Only 171 promoters show mutations in more than 5 samples per tumour type in a 200-bp window upstream of the promoter (Extended Data Fig. 15g, h). Most mutations occur in skin melanoma and lymphoma, which is expected owing to reduced nucleotide-excision repair and activation-induced cytidine deaminase (Extended Data Fig. 15h). We did not find significant pan-cancer associations between promoter mutational burden and promoter activity (Extended Data Fig. 15i, j). However, *TERT* has the highest number of promoter mutations<sup>1,5,31</sup> (Extended Data Fig. 16a), and these mutations have previously been reported to be associated with *TERT* expression<sup>1</sup>; therefore, we investigated the *TERT* locus in more detail (Extended Data Fig. 16b). Although *TERT* does not show a significant association in the



**Fig. 2 | Position-specific effect of somatic mutations on alternative splicing.** **a**, Top, proportion of mutations near exon–intron junctions and at branch sites that are associated with exon-skipping events. Mutations with associated splicing changes are those in which the percentage spliced in-derived |z-score| is  $\geq 3$  (dark blue). Asterisks denote intron positions significantly enriched for splicing changes relative to background based on a permutation test. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . Bottom, sequence motifs of regions. **b**, Example of an

exonization event in the tumour-suppressor gene *STK11*. The RNA-seq read coverage for a part of the gene is shown in red for a donor carrying the alternative allele, and in grey for a random donor with reference allele. The cassette exon event is shown as a schematic below. **c**, Enrichment of SINE elements in SAVs compared to sequence background (BG). Shown for SINE elements overlapping in sense (middle) and antisense (right) directions.

pan-cancer analysis, we found an association with increased promoter activity in individual types of cancer<sup>1</sup> (Extended Data Fig. 16c).

## Mutations associated with splicing

Extending the classical hallmarks of cancer, alternative splicing is seen as increasingly relevant to explain cancer heterogeneity<sup>32</sup>. On the basis of our observations of a globally changing splicing landscape (Extended Data Fig. 17a–c), we sought to specifically understand the relationship between splicing changes and somatic mutations within introns. Focusing on cassette exon events, we integrated the quantification of splice events with somatic variants and identified 5,282 mutations near exon–intron boundaries, 1,800 (34%) of which were associated with a change in splicing ( $|z\text{-score}| \geq 3$ ) (Supplementary Table 10). Consistent with previous findings using exome sequencing<sup>33,34</sup>, most mutations overlapping the essential dinucleotide motifs of the acceptor or donor site are associated with a splicing change—61% or 57%, respectively (Fig. 2a). Nearly one-third of all mutations (226 out of 469) in a 5-nucleotide window downstream of the 5' site were significantly enriched for splicing changes (Fig. 2a). Almost all changes significantly associated with somatic mutations had a negative effect on splicing (96%) (Extended Data Fig. 17d). For mutations in or near the poly-pyrimidine tract, we found a significant (permutation test,  $P < 0.05$ ) enrichment for mutations linked to outlier splicing (Fig. 2a). We also found an enrichment ( $P < 0.05$ , fold change  $> 2$ ) of splicing outliers at branch-site adenosines (Fig. 2a middle, Extended Data Fig. 17d, Supplementary Table 11). Together, these results suggest that somatic mutations in the extended splice site region, poly-pyrimidine tract and branch point can affect splicing.

We also identified 1,900 rare splicing-associated variants (SAVs) that appear in only a small number of samples using the SAVNet approach<sup>35</sup> (Extended Data Fig. 17e; see 'Data availability' in the Methods). Notably, 862 SAVs affected canonical splice sites, whereas the other 1,038 disrupted non-canonical sites or created new splice sites. Notably, we find a twofold enrichment of cancer genes in SAVs (Extended Data Fig. 17f).

Although we find that those SAVs that create splice sites strongly concentrate near exon–intron boundaries (Extended Data Fig. 17g), 45.9% of SAVs are further than 100 bp away from the nearest annotated exon. Mutations at those sites generally changed the sequences towards the donor or acceptor motif consensus (Extended Data Fig. 17h). Focusing on novel splice sites deep in introns, we analysed the extent of exonization—that is, the formation of new exons within an intron (Extended

Data Fig. 17j, Supplementary Tables 13, 14). More than one-fifth of these new exons (9 out of 43) occur in cancer-related genes, such as the well-known tumour-suppressor gene *STK11*. As expected, the exonization event would cause a frameshift in *STK11* (Fig. 2b, Extended Data Fig. 17k).

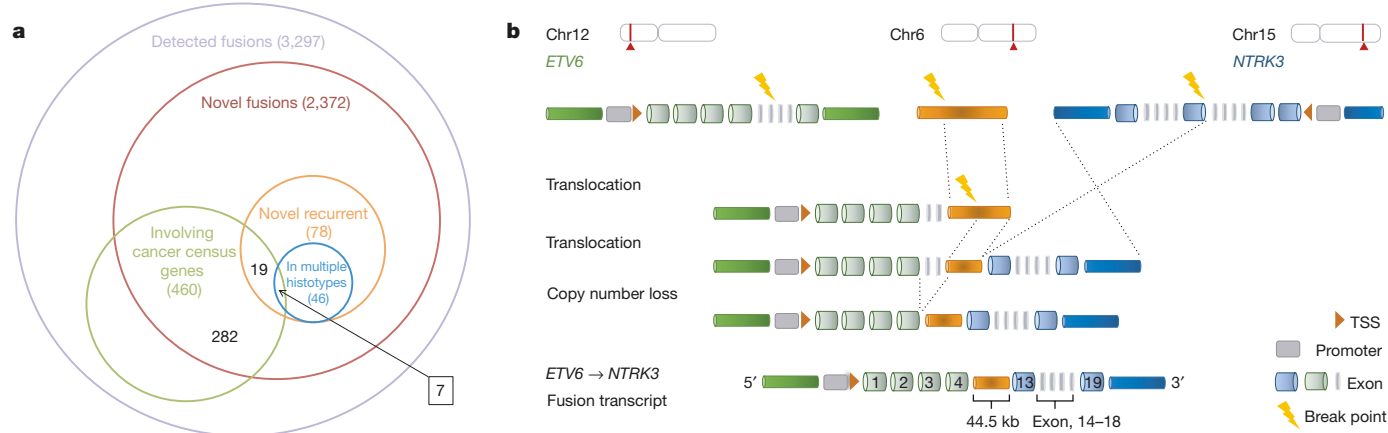
Alu elements that are inserted in an antisense direction have sequences that resemble consensus splice sites that, together with activating mutations, can lead to the formation of a new exon<sup>36</sup> (Extended Data Fig. 17l). We found a significant enrichment of splice-site-creating SAVs within annotated Alu sequences ( $P = 2.8 \times 10^{-9}$ ), particularly in the antisense direction ( $P = 2.6 \times 10^{-15}$ ) (Fig. 2c). Our results indicate that the exonization of Alu sequences, which has been extensively studied in the context of primate genome evolution, is also observed in cancer genome evolution.

## Patterns of gene fusions across cancer

Gene fusions are an important class of cancer-driving event with therapeutic and diagnostic value<sup>37</sup>. We identified a total of 925 known and 2,372 new cancer-specific gene fusions by combining the output of two fusion discovery methods as well as genomic rearrangement (structural variants) information and excluding artefacts or fusions in non-cancer samples<sup>38</sup> (Fig. 3a). For the 3,540 identified fusion events representing 3,297 unique gene fusions, we categorized them on the basis of novelty, recurrence and known oncogenic gene partners (Fig. 3a).

Only 149 (approximately 5%) of the fusions occur in more than one sample, among which 78 are novel. Most of these (46 out of 78) were found across several histotypes. Of the 27 most recurrent gene fusions (Extended Data Fig. 18a), 8 have previously been reported (for example, *CCDC6-RET*<sup>39</sup>, *FGFR3-TACC3*<sup>40</sup> and *PTPRK-RSPO3*) or independently detected in the TCGA cohort<sup>41</sup>, whereas 6 were new (such as *NUMB-HEATR4*, *ESR1-AKAP12* and *TRAF3IP2-FYN*). In total, 105 fusion transcripts involved the UTR region of one gene and the complete coding sequences of another gene, possibly resulting from structural variation in promoter regions.

Although most genes involved in fusions engaged with only one fusion partner, 35 genes had more than 5 partners. These 'promiscuous' genes tended to be selective in being either a 5' or a 3' partner with conserved break points and positions (3' or 5'), and were over-represented in cancer census genes and the PCAWG cancer-driver genes (one-tailed Fisher's exact test, odds ratio = 8.66,  $P \leq 1.1 \times 10^{-15}$ , and odds ratio = 12.27,  $P \leq 2.2 \times 10^{-16}$ , respectively). Network analysis of promiscuous genes and their partners revealed several large gene clusters containing at least 10 genes (Extended Data Fig. 18b), enriched



**Fig. 3 | Structural rearrangements associated with RNA fusions. a**, The number of all detected and new fusions and their overlap with the cancer census genes. **b**, Schematic of an example of bridged fusions. Bridged fusions

are those composite fusions formed by a third genomic segment that bridges two genes. Only one of the possible orders of genomic arrangement is depicted in each case, with break points highlighted as thunderbolts.

in cancer-related pathways (Benjamini–Hochberg corrected  $P \leq 0.01$ ) and in protein–protein interactions ( $P \leq 1.0 \times 10^{-7}$ ), which suggests a possible functional role in cancer.

Notably, a large number of fusions, including known fusions, could not be associated with only a single structural-variation event. For example, the *ETV6-NTRK3* gene fusion<sup>42</sup> was present in a head and neck thyroid carcinoma sample, linking exon 4 of *ETV6* to exon 12 of *NTRK3*. We found three separate structural variants in the same sample: (1) a translocation of *ETV6* to chromosome 6; (2) a translocation of *NTRK3* also to chromosome 6; and (3) an additional copy-number loss spanning from intron 5 of *ETV6* to the exact structural variant break points, jointly bringing *ETV6* within 45 kb upstream of *NTRK3*—a distance that would allow transcriptional read-through<sup>43</sup> or splicing<sup>44</sup> to yield the *ETV6-NTRK3* fusion<sup>45</sup> (Fig. 3b). Thus, the short chromosome-6 segment appeared to function as a bridge, which linked two genomic locations to facilitate a gene fusion. We term such products ‘bridged fusions’. This class of fusion is not uncommon. Out of a total of 436 gene fusions supported by 2 separate structural variants, 75 are bridged fusions (Supplementary Table 15).

On the basis of the nature of the underlying genomic rearrangements, we propose a unified fusion classification system (Extended Data Fig. 19a). Aside from bridged fusions, 344 additional fusions are linked to more than one structural variant in the same sample. These multi-structural variant fusions are collectively termed ‘composite fusions’ (Extended Data Fig. 19a, b). We find 284 intercomposite fusions (interchromosomal translocation) and 124 intracomposite fusions (intrachromosomal rearrangement), exemplified by *ERC1-RET1* and *NUMB-HEATR4* fusions, respectively (Extended Data Fig. 19b). Composite rearrangements bring the fusion partners significantly closer to each other, from the median natural distance of 6.8 Mb to the median of 7.9 kb (Wilcoxon rank-sum test,  $P \leq 2.2 \times 10^{-16}$ ; Extended Data Fig. 19c) after translocation. For 18% of fusions, no evidence of structural variation was found. Given that 340 structural-variant-independent, intrachromosomal fusions had significantly closer break points than those with structural variation (Extended Data Fig. 19d), it is possible that they could result from RNA read-through events. The other possibility is that the underlying supporting structural variants escaped detection, as shown by the observation that known gene fusions that are driven by structural variation, such as *TMPRSS2-ERG*<sup>46</sup>, did not have consistent evidence for structural variation in matching samples.

## Landscape of RNA alterations in cancer

Given our comprehensive set of RNA alterations, we sought to characterize the heterogeneous mechanisms of cancer genome and

transcriptome alterations. To enable joint analyses of RNA and DNA alterations, we created a gene-level table, which indicates the presence or absence of possible functional changes to RNA or DNA for each gene and donor. After stringent filtering, we identified 1,523,098 alteration events, in which an event is a gene–sample–alteration triplet (Extended Data Table 1, Supplementary Table 14). It should be noted that we chose to include only RNA alterations with potential functional effects or with the strongest quantitative affect, resembling similar strategies for filtering DNA alterations<sup>47</sup>. Recurrence analysis across several alteration types helped us to further enrich for functionally relevant genes. Building on the gene-centric table, we characterized gene alterations at the RNA level and contrasted these with DNA alterations (non-synonymous SNVs or SCNAs)<sup>5</sup>. On the basis of the calculated association between each RNA- and DNA-level alteration across all histotypes, we found that half of the RNA alterations significantly correlated with DNA alterations (likelihood ratio test,  $FDR < 1 \times 10^{-4}$ ) (Extended Data Fig. 20).

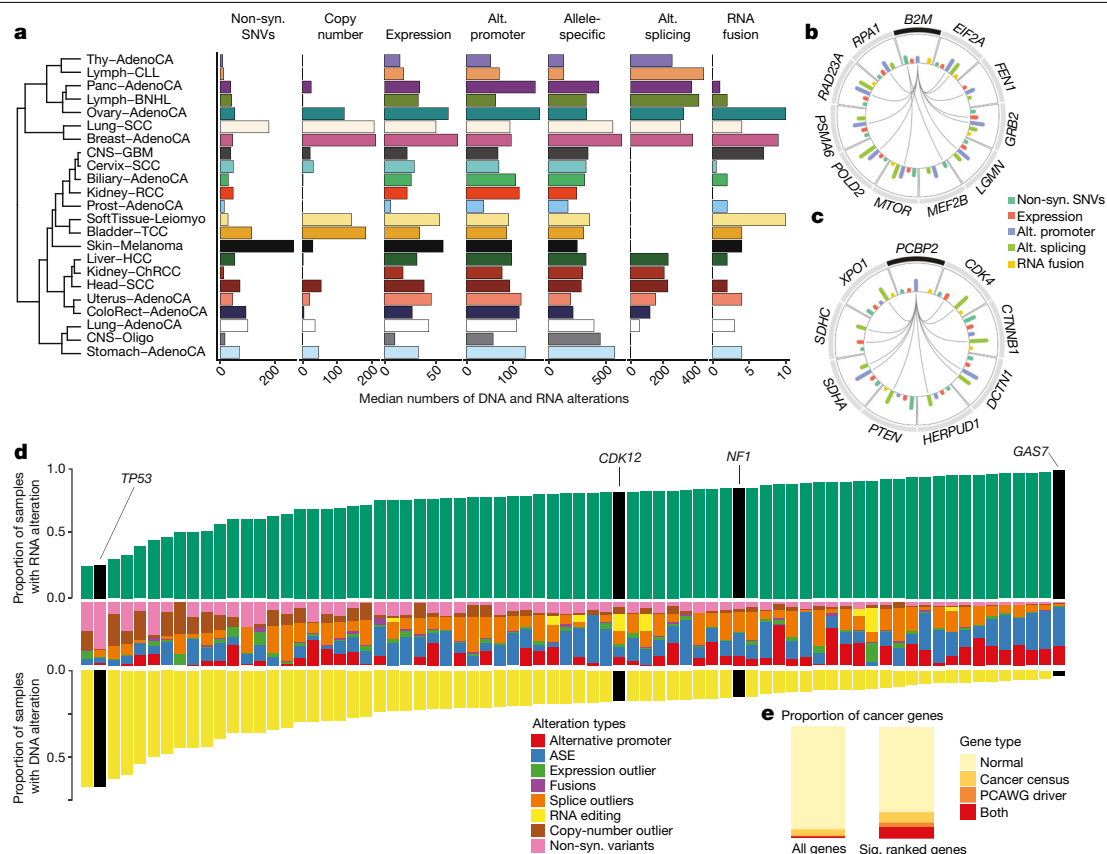
When comparing gene alteration frequencies across all histotypes (Fig. 4a), we note that different types of cancer contain distinct combinations of DNA- and RNA-level alterations (Fig. 4a, Supplementary Table 17). Although, as expected, skin melanoma significantly exceeds other cancers in the number of non-synonymous SNVs<sup>48</sup> (Wilcoxon rank-sum test,  $P < 0.012$ ), lymphatic cancers have low numbers of SNVs (Wilcoxon rank-sum test,  $P = 5.3 \times 10^{-15}$ ), but high incidences of alternative splicing outliers (Wilcoxon rank-sum test,  $P = 4.9 \times 10^{-47}$ ), which suggests that transcriptomic alterations can be relatively more pronounced in certain cancer types.

To evaluate to which extent RNA changes provide additional mechanisms for cancer gene alterations, we examined DNA- and RNA-level alterations both in sets of genes in pathways (Extended Data Fig. 21) and in individual genes with known roles in cancer (Extended Data Fig. 22). We found that RNA alterations occur at a high proportion in many pathways, including the NOTCH and TGF- $\beta$  pathways. In addition, *KRAS* exhibits more RNA alterations than DNA alterations in some types of cancer. Given the recent finding that alternative splicing of *KRAS* expanded the prognostic affect beyond mutation status in colorectal cancer<sup>49</sup>, our data further support several modes of alteration for *KRAS* in tumours.

## Co-occurrence of RNA and DNA alterations

The diverse types of alteration in this study enabled us to investigate *trans*-associations between different genetic and expression characteristics involving cancer-related genes ( $FDR < 5\%$ ) (Supplementary Table 18). By investigating whether somatic mutations of known cancer





**Fig. 4 | Global view of DNA and RNA alterations that affect tumours. a**, The median numbers of different alterations across histotypes. Histotypes are ordered by hierarchical clustering based on the pattern of different types of alteration. Only histotypes with more than 10 donors are shown. Alt., alternative; non-syn, non-synonymous. Cancer-type abbreviations are listed in Supplementary Table 23. **b, c**, Circular representations of the selected genes significantly co-occurred with *B2M* (**b**) and *PCBP2* (**c**). Connecting lines indicate the specific types of co-occurrence of alteration pairs. The inner histograms

indicate the frequencies of incidences of different alteration types shown in different colours. **d**, All 74 Catalogue of Somatic Mutations in Cancer (COSMIC) cancer census genes or PCAWG driver genes that are both frequently and heterogeneously altered across both RNA- and DNA-level alterations. Yellow bars indicate the proportion of samples that had DNA-level alterations, and green bars indicate the proportion of samples with RNA-level alterations. Middle column is the proportion of each alteration type observed for that gene. **e**, The enrichment of cancer genes within our list of significantly recurrent genes.

genes are associated with the expression of other genes, we found *IDH1* and *NFKBIE* to be widely linked to the dysregulation of many genes (Extended Data Fig. 23a, b). Notable co-occurrences were present in several types of cancer. For example, *B2M* and *EIF4G2* alterations were simultaneously observed in both B-cell non-Hodgkin lymphoma and lung squamous cell carcinoma. Pathway enrichment analysis of the top 100 genes associated with all *B2M* alterations indicates that the most affected genes are involved in DNA repair ( $FDR \leq 1\%$ ), and approximately two-thirds of those associations were significant in more than one cancer type (Fig. 4b, Extended Data Fig. 23c).

We also examined how cancer genes could be affected by other genes by co-occurrence analyses. Expression outliers of *PCBP2* co-occurred with aberrant splicing of a large number of cancer-related genes, including *CTNMB1* and *CDK4* (Fig. 4c). *PCBP2* has been reported to enhance the splicing of cassette exons<sup>50</sup>. Our results thus further support the possible role of *PCBP2* in regulating the splicing of cancer-related genes.

### Recurrent RNA alterations in driver genes

In our analyses of *cis*-acting mutations that are associated with these individual RNA phenotypes, the vast majority were observed rarely in the PCAWG cohort. Many cancer genes (such as *MET*<sup>51,52</sup>) are known to be somatically altered by heterogeneous mechanisms such as gene fusions, splicing mutations and non-synonymous mutations; therefore, examining genes that are altered by several *cis*-acting mechanisms may help to identify cancer genes in which an individual alteration

type is rare. A total of 5,413 genes were altered by gene expression, allele-specific expression (ASE), splicing and/or gene fusion, and had an associated DNA-level mutation in *cis* (Supplementary Table 20). PCAWG-defined driver genes<sup>8</sup> tended to have more diverse mechanisms of RNA-level alterations when compared to genes that have not previously been identified as a cancer gene ( $P < 0.001$ ) (Extended Data Fig. 24a). We identified, for example, a somatic eQTL, a splicing-associated variant and fusions in the known tumour-suppressor *NF1* in the MAPK pathway (Extended Data Fig. 24b).

Owing to the fact that most somatic mutations are rare<sup>5</sup>, it is difficult to statistically distinguish functionally relevant, potential driver alterations from passenger alterations. Therefore, we aimed to identify genes that are both recurrently and heterogeneously altered, under the hypothesis that these genes have increased functional relevance. This analysis identified 731 genes with significant recurrent aberrations ( $FDR < 5\%$ ) (Extended Data Fig. 25a), with the top-ranking genes carrying both RNA and DNA alterations. RNA alterations accounted for 0.05–99.14% (mean: 78.23%) of all identified alterations in each gene (Extended Data Fig. 25a, Supplementary Table 21). This ranking is enriched for the union of cancer census genes<sup>53</sup> (60 out of 603) and PCAWG-defined driver genes (33 out of 157, unioned: 74 out of 674  $P = 4.6 \times 10^{-13}$ , enrichment: 2.45) (Fig. 4d, e).

Among the top 10% of our ranked genes is *CDK12* (rank 55). We find 91 samples that have an alteration involving its protein kinase domain, which has been implicated in DNA repair dysregulation<sup>54</sup>. Many of these samples have no DNA-level alterations in *CDK12* (46%) (Extended Data



Fig. 26a). Furthermore, splicing, alternative promoter, SNV, RNA-editing and fusion alterations in this gene are mutually exclusive (adjusted  $P=4.8 \times 10^{-3}$ ) (Extended Data Fig. 26b, c). Upon further investigation, we find that somatic eQTL mutations in *CDK12* are associated with a tandem duplicator phenotype<sup>55</sup>. Although this association was not replicated with other RNA alterations, it provides evidence that somatic *CDK12* mutations may alter its function through gene expression changes. This example illustrates that performing a recurrence analysis over diverse RNA and DNA alterations can help to identify genes known to be important in tumorigenesis.

## Discussion

Here we present a comprehensive catalogue of RNA-level alterations in cancer, spanning 27 different tumour types, and provide a harmonized resource of matched transcriptome and whole-genome sequences. We identified 731 genes that were recurrently altered by several mechanisms, jointly enriched for known cancer census and PCAWG driver genes<sup>8</sup>. The list includes genes that are primarily altered at the DNA level (such as *TP53*), but also genes for which the alteration most frequently manifests in RNA (such as *GAS7*). Out of 87 samples from the PCAWG study that did not have a driver alteration at the DNA level<sup>5</sup>, and had RNA-sequencing (RNA-seq) data, every sample had an RNA-level alteration identified. Although cancer is thought to be driven by changes in DNA primarily, some driver alterations may manifest themselves via changes in RNA rather than DNA sequence mutations.

We identified germline eQTLs for around 20% of expressed genes. The number of eGenes found is generally low compared with some other studies, reflecting the heterogeneity of our samples. Only 422 genes appeared to be specific to cancer; this is likely to be an underestimate owing to the heterogeneity, small sample numbers and the rather conservative strategy chosen. We have also mapped linkages between genes and somatic aberrations in *cis*, in which 68.4% of associations were between non-coding somatic variants and gene expression. Allelic copy-number imbalance is a major determinant of ASE dysregulation in cancer. We found mutations associated with splicing changes including novel cancer-specific exons that can be partially explained by mutation-driven exonization. We systematically compared gene fusions with whole-genome rearrangements across many tumour types and found 82% of detected fusions were associated with specific genomic rearrangements. For the remaining fusions, it is possible that the relevant genomic rearrangements have not been detected, or that some fusions happen directly at the RNA level, as *trans*-splicing or read-through events. The availability of whole-genome sequences allowed us to develop a systematic classification of fusion events and to propose a new bridged fusion mechanism.

Because global differences in RNA expression phenotypes are largely tissue-specific, our ability to associate mutations in *cis* or *trans* are limited by the small and variable sample sizes within each histotype. Further work is needed to investigate other mechanisms of genome alteration that can lead to changes in RNA such as epigenetic changes<sup>56</sup> or enhancer hijacking<sup>57</sup>. Our work will help to prioritize further investigations.

Overall, our analyses show diverse modes of alteration of cancer genes and pathways at the DNA and RNA levels, and demonstrate that RNA analyses reveal cancer-associated pathway alterations that have not yet been detected via DNA-only approaches. These insights illustrate the power of integrated transcriptome and whole-genome sequencing analysis for cancer studies.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1970-0>.

- Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
- Owens, M. A., Horten, B. C. & Da Silva, M. M. HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. *Clin. Breast Cancer* **5**, 63–69 (2004).
- Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyraes, E. The functional impact of alternative splicing in cancer. *Cell Reports* **20**, 2215–2226 (2017).
- Faderl, S. et al. The biology of chronic myeloid leukemia. *N. Engl. J. Med.* **341**, 164–172 (1999).
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* <https://doi.org/10.1038/s41586-019-1907-7> (2020).
- Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* <https://doi.org/10.1038/s41586-019-1913-9> (2020).
- Rheinbay, E. et al. Analyses of non-coding somatic mutations in 2,693 cancer whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1965-x> (2020).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* <https://doi.org/10.1038/s41586-020-1943-3> (2020).
- GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
- Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
- Gong, J. et al. PanCanQTL: systematic identification of *cis*-eQTLs and *trans*-eQTLs in 33 cancer types. *Nucleic Acids Res.* **46**, D971–D976 (2018).
- Bajrami, I. et al. Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res.* **74**, 287–297 (2014).
- Havelange, V. et al. IRF4 mutations in chronic lymphocytic leukemia. *Blood* **118**, 2827–2829 (2011).
- Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Zheng, C. L. et al. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Reports* **9**, 1228–1234 (2014).
- Hanafusa, T., Mohamed, A. E. A., Domae, S., Nakayama, E. & Ono, T. Serological identification of Tekin5 as a cancer/testis antigen and its immunogenicity. *BMC Cancer* **12**, 520 (2012).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- Milacic, M. et al. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel)* **4**, 1180–1211 (2012).
- Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).
- Kvam, E. & Tyrrell, R. M. The role of melanin in the induction of oxidative DNA base damage by ultraviolet A irradiation of DNA or melanoma cells. *J. Invest. Dermatol.* **113**, 209–213 (1999).
- Jimbow, K., Chen, H., Park, J. S. & Thomas, P. D. Increased sensitivity of melanocytes to oxidative stress and abnormal expression of tyrosinase-related protein in vitiligo. *Br. J. Dermatol.* **144**, 55–65 (2001).
- Pilger, A. & Rüdiger, H. W. 8-Hydroxy-2'-deoxyguanosine as a marker of oxidative DNA damage related to occupational and environmental exposures. *Int. Arch. Occup. Environ. Health* **80**, 1–15 (2006).
- Premi, S. & Brash, D. E. Unanticipated role of melanin in causing carcinogenic cyclobutane pyrimidine dimers. *Mol. Cell. Oncol.* **3**, e1033588 (2015).
- Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21**, 457–465 (2005).
- Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
- Demircioğlu, D. et al. A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. *Cell* **178**, 1465–1477.e17 (2019).
- Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018).
- Feng, G. et al. Ubiquitously expressed genes participate in cell-specific functions via alternative promoter usage. *EMBO Rep.* **17**, 1304–1313 (2016).
- Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
- Oltean, S. & Bates, D. O. Hallmarks of alternative splicing in cancer. *Oncogene* **33**, 5311–5318 (2014).
- Jung, H. et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).
- Kahles, A. et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* **34**, 211–224.e6 (2018).
- Shiraishi, Y. et al. A comprehensive characterization of *cis*-acting splicing-associated variants in human cancer. *Genome Res.* **28**, 1111–1125 (2018).
- Sorek, R. The birth of new exons: mechanisms and evolutionary consequences. *RNA* **13**, 1603–1608 (2007).
- Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
- Mélè, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

39. Matsubara, D. et al. Identification of *CCDC6-RET* fusion in the human lung adenocarcinoma cell line, LC-2/ad. *J. Thorac. Oncol.* **7**, 1872–1876 (2012).
40. Carneiro, B. A. et al. *FGFR3-TACC3*: a novel gene fusion in cervical cancer. *Gynecol Oncol Rep* **13**, 53–56 (2015).
41. Lee, M. et al. ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res.* **45** (D1), D784–D789 (2017).
42. Knezevich, S. R., McFadden, D. E., Tao, W., Lim, J. F. & Sorensen, P. H. A novel *ETV6-NTRK3* gene fusion in congenital fibrosarcoma. *Nat. Genet.* **18**, 184–187 (1998).
43. Nacu, S. et al. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genomics* **4**, 11 (2011).
44. Jia, Y., Xie, Z. & Li, H. Intergenic spliced chimeric RNAs in cancer. *Trends Cancer* **2**, 475–484 (2016).
45. Greger, L. et al. Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. *PLoS ONE* **9**, e104567 (2014).
46. Tomlins, S. A. et al. Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
47. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
48. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
49. Eilertsen, I. A. et al. Alternative splicing expands the prognostic impact of *KRAS* in microsatellite stable primary colorectal cancer. *Int. J. Cancer* **144**, 841–847 (2019).
50. Ji, X. et al.  $\alpha$ CP binding to a cytosine-rich subset of polypyrimidine tracts drives a novel pathway of cassette exon splicing in the mammalian transcriptome. *Nucleic Acids Res.* **44**, 2283–2297 (2016).
51. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat. Commun.* **5**, 4846 (2014).
52. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
53. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45** (D1), D777–D783 (2017).
54. Blazek, D. et al. The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* **25**, 2158–2172 (2011).
55. Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. *Cancer Cell* **34**, 197–210.e5 (2018).
56. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150**, 12–27 (2012).
57. Zhang, X. et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.* **48**, 176–182 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

<sup>1</sup>A list of members and their affiliations appears at the end of the paper. <sup>2</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. <sup>3</sup>ETH Zurich, Zurich, Switzerland. <sup>4</sup>Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>5</sup>Weill Cornell Medical College, New York, NY, USA. <sup>6</sup>SIB Swiss Institute of Bioinformatics, Lausanne,

Switzerland. <sup>7</sup>University Hospital Zurich, Zurich, Switzerland. <sup>8</sup>National University of Singapore, Singapore, Singapore. <sup>9</sup>Genome Institute of Singapore, Singapore, Singapore. <sup>10</sup>Peking University, Beijing, China. <sup>11</sup>The University of Tokyo, Minato-ku, Japan. <sup>12</sup>University of California, Santa Cruz, Santa Cruz, CA, USA. <sup>13</sup>BGI-Shenzhen, Shenzhen, China. <sup>14</sup>China National GeneBank-Shenzhen, Shenzhen, China. <sup>15</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>16</sup>University of California, San Francisco, San Francisco, CA, USA. <sup>17</sup>University of Glasgow, Glasgow, UK. <sup>18</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. <sup>19</sup>The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>20</sup>Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. <sup>21</sup>University College London, London, UK. <sup>22</sup>Karolinska Institutet, Stockholm, Sweden. <sup>23</sup>Broad Institute, Cambridge, MA, USA. <sup>24</sup>Ulm University and Ulm University Medical Center, Ulm, Germany. <sup>25</sup>Duke-NUS Medical School, Singapore, Singapore. <sup>26</sup>University of Toronto, Toronto, Ontario, Canada. <sup>27</sup>Baylor College of Medicine, Houston, TX, USA. <sup>28</sup>Dana-Farber Cancer Institute, Boston, MA, USA. <sup>29</sup>Harvard Medical School, Boston, MA, USA. <sup>30</sup>University of Toronto, Toronto, Ontario, Canada. <sup>31</sup>National Cancer Centre Singapore, Singapore, Singapore. <sup>32</sup>German Cancer Consortium (DKTK), partner site Berlin, Germany. <sup>33</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>34</sup>A list of members and their affiliations appears in the Supplementary Information. <sup>35</sup>These authors contributed equally: PCAWG Transcriptome Core Group, Claudia Calabrese, Natalie R. Davidson, Deniz Demircioğlu, Nuno A. Fonseca, Yao He, André Kahles, Kjong-Van Lehmann, Fenglin Liu, Yuichi Shiraishi, Cameron M. Soulette, Lara Urban. <sup>36</sup>These authors jointly supervised this work: Alvis Brazma, Angela N. Brooks, Jonathan Göke, Gunnar Rättsch, Roland F. Schwarz, Oliver Stegle, Zemin Zhang. \*e-mail: [brazma@ebi.ac.uk](mailto:brazma@ebi.ac.uk); [anbrooks@ucsc.edu](mailto:anbrooks@ucsc.edu); [raetsch@inf.ethz.ch](mailto:raetsch@inf.ethz.ch)

## PCAWG Transcriptome Core Group

Claudia Calabrese<sup>2</sup>, Natalie R. Davidson<sup>3,4,5,6,7</sup>, Deniz Demircioğlu<sup>8,9</sup>, Nuno A. Fonseca<sup>2</sup>, Yao He<sup>10</sup>, André Kahles<sup>3,4,6,7</sup>, Kjong-Van Lehmann<sup>3,4,6,7</sup>, Fenglin Liu<sup>10</sup>, Yuichi Shiraishi<sup>11</sup>, Cameron M. Soulette<sup>12</sup> & Lara Urban<sup>2</sup>

## PCAWG Transcriptome Working Group

Nuno A. Fonseca<sup>2</sup>, André Kahles<sup>3,4,6,7</sup>, Kjong-Van Lehmann<sup>3,4,6,7</sup>, Lara Urban<sup>2</sup>, Cameron M. Soulette<sup>12</sup>, Yuichi Shiraishi<sup>11</sup>, Fenglin Liu<sup>10</sup>, Yao He<sup>10</sup>, Deniz Demircioğlu<sup>8,9</sup>, Natalie R. Davidson<sup>3,4,5,6,7</sup>, Claudia Calabrese<sup>2</sup>, Junjun Zhang<sup>15</sup>, Marc D. Perry<sup>15,16</sup>, Qian Xiang<sup>15</sup>, Liliana Greger<sup>2</sup>, Siliang Li<sup>13,14</sup>, Dongbing Liu<sup>13,14</sup>, Stefan G. Stark<sup>3,4,6,7</sup>, Fan Zhang<sup>10</sup>, Samirkumar B. Amin<sup>37</sup>, Peter Bailey<sup>17</sup>, Aurélien Chateigner<sup>15</sup>, Isidro Cortés-Ciriano<sup>29,38,39</sup>, Brian Craft<sup>12</sup>, Serap Erkek<sup>18</sup>, Milana Frenkel-Morgenstern<sup>40</sup>, Mary Goldman<sup>12</sup>, Katherine A. Hoadley<sup>19</sup>, Yong Hou<sup>13,14</sup>, Matthew R. Huska<sup>20</sup>, Ekta Khurana<sup>5</sup>, Helena Kilpinen<sup>21</sup>, Jan O. Korbel<sup>18</sup>, Fabien C. Lamaze<sup>15</sup>, Chang Li<sup>13,14</sup>, Xiaobo Li<sup>13,14</sup>, Xinyue Li<sup>13</sup>, Xingmin Liu<sup>13,14</sup>, Maximilian G. Marin<sup>12</sup>, Julia Markowski<sup>20</sup>, Tannistha Nandi<sup>9</sup>, Morten M. Nielsen<sup>41</sup>, Akinyemi I. Ojesina<sup>23,28,42,43</sup>, Qiang Pan-Hammarström<sup>13,22</sup>, Peter J. Park<sup>29,38</sup>, Chandra Sekhar Pedamallu<sup>23,28,29</sup>, Jakob S. Pedersen<sup>41</sup>, Reiner Siebert<sup>24</sup>, Hong Su<sup>13,14</sup>, Patrick Tan<sup>9,25</sup>, Bin Tean Teh<sup>31</sup>, Jian Wang<sup>13</sup>, Sebastian M. Waszak<sup>18</sup>, Heng Xiong<sup>13,14</sup>, Sergei Yakneen<sup>18</sup>, Chen Ye<sup>13,14</sup>, Christina Yung<sup>15</sup>, Xiuqing Zhang<sup>13</sup>, Liangtao Zheng<sup>10</sup>, Jingchun Zhu<sup>12</sup>, Shida Zhu<sup>13,14</sup>, Philip Awadalla<sup>15,26</sup>, Chad J. Creighton<sup>27</sup>, Matthew Meyerson<sup>23,28,29</sup>, B. F. Francis Ouellette<sup>30</sup>, Kui Wu<sup>13,14</sup>, Huanming Yang<sup>13</sup>, Jonathan Göke<sup>9,31</sup>, Roland F. Schwarz<sup>2,20,32,33</sup>, Oliver Stegle<sup>2,18,33</sup>, Zemin Zhang<sup>10</sup>, Alvis Brazma<sup>2</sup>, Gunnar Rättsch<sup>3,4,5,6,7</sup> & Angela N. Brooks<sup>12,23,28</sup>

<sup>37</sup>The UT MD Anderson Cancer Center, Houston, TX, USA. <sup>38</sup>Ludwig Center at Harvard, Boston, MA, USA. <sup>39</sup>University of Cambridge, Cambridge, UK. <sup>40</sup>The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. <sup>41</sup>Aarhus University, Aarhus, Denmark. <sup>42</sup>HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. <sup>43</sup>University of Alabama at Birmingham, Birmingham, AL, USA.

## Methods

### RNA-seq alignment and quality-control analysis

Tumour and healthy ICGC RNA-seq data, included in the PCAWG cohort<sup>5</sup>, was aligned to the human reference genome (GRCh37.p13) using two read aligners: STAR<sup>58</sup> (v.2.4.0i, two-pass), performed at MSKCC and ETH Zürich, and TopHat2<sup>59</sup> (v.2.0.12), performed at the European Bioinformatics Institute. Both tools used Gencode (release 19)<sup>60</sup> as the reference gene annotation. For the STAR two-pass alignment, an initial alignment run was performed on each sample to generate a list of splice junctions derived from the RNA-seq data. These junctions were then used to build an augmented index of the reference genome per sample. In a second pass, the augmented index was used for a more sensitive alignment. Alignment parameters have been fixed to the values reported in <https://github.com/ICGC-TCGA-PanCancer/pcawg3-rnaseq-align-star>. The TopHat2 alignment strategies also followed the two-pass alignment principle, but was performed in a single alignment step with the respective parameter set. For the TopHat2 alignments, the irap analysis suite<sup>61</sup> was used. The full set of parameters is available along with the alignment code in [https://hub.docker.com/r/nunofonseca/irap\\_pcawg/](https://hub.docker.com/r/nunofonseca/irap_pcawg/). For both aligners, the resulting files in BAM format were sorted by alignment position, indexed and are available for download in the GDC portal (<https://portal.gdc.cancer.gov/>) and the ICGC Data Portal (<https://dcc.icgc.org/>). The individual accession numbers and download links can be found in the PCAWG data release table: [http://pancancer.info/data\\_releases/may2016/release\\_may2016.v1.4.tsv](http://pancancer.info/data_releases/may2016/release_may2016.v1.4.tsv). Cancer-type abbreviations are listed in Supplementary Table 23. Histology was derived from an older version released by the PCAWG Pathology and Clinical Correlates Working Group. Assignments of donor to histology used in this study can be found in the file `rnaseq.extended.metadata.aliquot_id.V4.tsv.gz` at <https://dcc.icgc.org/releases/PCAWG/transcriptome/metadata/>.

Quality control of all datasets was performed at three main levels: (1) assessment of initial raw data using FastQC<sup>62</sup> (v.0.11.3) (Supplementary Fig. 4); (2) assessment of aligned data (percentage of mapped and unmapped reads for both alignment approaches); and (3) quantification (by correlating the expression values produced by the STAR and TopHat2 based expression pipelines) (Supplementary Fig. 2). In total, we defined six quality-control criteria to assess the quality of the samples. We marked a sample as a candidate for exclusion if: (1) 3 out of 5 main FastQC measures (base-wise quality, *k*-mer overrepresentation, guanine-cytosine content, content of *N* bases and sequence quality) did not pass; (2) more than 50% of reads were unmapped or fewer than 1 million reads could be mapped in total using the STAR pipeline; (3) more than 50% of reads were unmapped or fewer than 1 million reads could be mapped in total using the TopHat2 pipeline; (4) we measured a degradation score<sup>63</sup> greater than 10; (5) the fragment count in the aligned sample (averaged over STAR and TopHat2) was <5 million; and (6) the correlation between the expression counts of both pipelines was <0.95. If a sample did not pass one of these six criteria it was marked as problematic and placed on a greylist. If more than two criteria were not passed, we excluded the sample.

A subset of 722 libraries from the projects ESAD-UK, OV-AU, PACA-AU and STAD-US were identified as technical replicates generated from the same sample aliquot. These libraries were integrated post-alignment for both the STAR and the TopHat2 pipelines using samtools<sup>64</sup> into combined alignment files. Further analysis was based on these files. Read counts of the individual libraries were integrated to a sample-level count by adding the read counts of the technical replicates.

Initially, a total of 2,217 RNA-seq libraries were fully processed by the pipeline. Quality-control filtering and integration of technical replicates (722 libraries) gave a final number of 1,359 fully processed RNA-seq sample aliquots from 1,188 donors.

### GTEX data analysis

For a panel of RNA-seq data from a variety of healthy tissues, data from 3,274 samples from GTEx (phs000424.v4.p1) were used and analysed with the same pipeline as PCAWG data for quantifying gene expression. A list of GTEx identifiers are provided at <https://dcc.icgc.org/releases/PCAWG/transcriptome/metadata>.

### Quantification and normalization of transcript and gene expression

STAR and TopHat2 alignments were used as input for HTSeq<sup>65</sup> (v.0.6.1p1) to produce gene expression counts. Gencode v.19<sup>60</sup> was used as the gene annotation reference. Quantification on a per-transcript level was performed with Kallisto<sup>66</sup> (v.0.42.1). This implementation is available as a Docker container at [https://hub.docker.com/r/nunofonseca/irap\\_pcawg](https://hub.docker.com/r/nunofonseca/irap_pcawg). The implementation of the STAR and TopHat2 quantification is available as docker containers in: <https://github.com/ICGC-TCGA-PanCancer/pcawg3-rnaseq-align-star> and [https://hub.docker.com/r/nunofonseca/irap\\_pcawg/](https://hub.docker.com/r/nunofonseca/irap_pcawg/), respectively. Quantification of consensus expression was performed by taking the average expression based on STAR and TopHat2 alignments. Gene counts were normalized by adjusting the counts to FPKM<sup>67</sup> as well as FPKM with upper quartile normalization (FPKM-UQ) in which the total read counts in the FPKM definition has been replaced by the upper quartile of the read count distribution multiplied by the total number of protein-coding genes.

The FPKM and FPKM-UQ calculations were as follows.  $FPKM = (C \times 10^9) / (NL)$ , in which *N* denotes the total fragment count to protein-coding genes, *L* denotes the length of the gene and *C* denotes the fragment count.  $FPKM-UQ = (C \times 10^9) / (ULG)$ , in which *U* denotes the upper quartile of fragment counts to protein-coding genes on autosomes unequal to zero, and *G* denotes the number of protein-coding genes on autosomes.

### *t*-Distributed stochastic neighbour embedding analysis

The *t*-distributed stochastic neighbour embedding (*t*-SNE) plots in Supplementary Figs. 5 and 6 were produced using the RTsne package<sup>68</sup> (with a perplexity value of 3) based on the Pearson correlation of the aggregated expression ( $\log + 1$ ) of the 1,500 most variable genes. FPKM expression values per gene were aggregated (median) by tissue (GTEx) and study (PCAWG). Coefficient of variation for each gene was also computed per tissue (GTEx) and study (PCAWG) to determine the 1,500 most variable genes. Purity values were previously described<sup>69</sup>.

The *t*-SNE plot in Extended Data Fig. 17c is based on all exon-skipping events in protein-coding genes confirmed by SplAdder<sup>70</sup>. Each event was quantified in both the PCAWG and GTEx cohort. All events with more than 1% of missing percentage spliced in (PSI) values across the concatenated PCAWG and GTEx samples were removed. The remaining missing values were imputed as the mean over the non-missing samples. The centred data were then visualized using the TSNE package from the Scikit Learn toolkit<sup>71</sup> with a perplexity value of 100, random state 0 and an initialization with PCA.

### Associations between genetic variation and gene expression: patient cohort

To associate genetic variation with gene expression, we analysed whole-genome sequencing (WGS) of the 1,188 donors with matched whitelisted RNA-seq data from the PCAWG cohort. Germline genotypes, SNV calls and segmented allele-specific SCNA calls were previously reported<sup>5</sup>. We matched 1,188 tumour RNA-seq IDs<sup>5</sup> to WGS whitelist tumour IDs (synapse entry syn10389164). For patients with multiple WGS IDs (2 out of 1,188) or RNA-seq aliquot IDs (17 out of 1,188), we resolved the matching by pairing samples with the same 'tumor\_wgs\_submitter\_specimen\_id' (Supplementary Table 1). The 1,188 patients are spread across 27 types of cancer and 29 project codes and include

# Article

899 carcinomas; 34 patients are metastatic and 13 recurrent with the remaining patients being primary tumours (Supplementary Table 1).

We used the data of these 1,188 patients for performing somatic and germline eQTL mapping, ASE analysis and association studies between gene expression and mutational signatures.

## Gene expression filtering

Gene expression values (measured in FPKM; [https://dcc.icgc.org/releases/PCAWG/transcriptome/gene\\_expression](https://dcc.icgc.org/releases/PCAWG/transcriptome/gene_expression)) from consensus expression quantification as described above were used for this analysis.

Genes with FPKM  $\geq 0.1$  in at least 1% of the patients (12 patients) were retained, resulting in 47,730 genes. Only 18,898 protein-coding genes (according to the 'gene\_type' biotype reported in Gencode v.19<sup>60</sup>) were used for the subsequent QTL analyses. The  $\log_2$ -transformed expression values (FPKM + 1) were subjected to peer analysis<sup>72</sup> to account for hidden covariates (syn7850427; <https://dcc.icgc.org/releases/PCAWG/transcriptome/eQTL/phenotype>). To balance the number of covariates, statistical power and available sample sizes per cancer type, we followed the GTEx protocol and estimated 15, 30 and 35 hidden covariates to be used depending on sample size<sup>73</sup> ( $n < 150$ ,  $150 \leq n < 250$ ,  $n \geq 250$ ). Peer residuals were then rank-standardized across patients. The FPKM cut-off values and peer correction were also applied to the subset of 899 patients with carcinoma, yielding 18,837 protein-coding genes after filtering. Furthermore, we used ordinary least-squares regression to correlate each of the 35 peer factors with per-sample covariates, including cancer project codes, gender, tumour purity, somatic burden and several sequence metrics (Supplementary Notes), to understand the proportion of variance explained by known biological and technical covariates.

## Covariates

In all linear models, we accounted for known confounding factors by modelling them as fixed effects. In all association studies, we accounted for sex, project code (describing cancer type and country of origin) and per-gene copy-number status (Supplementary Table 1 for the list of per patient covariates; syn7253568 and syn7253569 for sex and project codes; syn9661460 for per gene copy number). Per-gene copy-number alterations were derived as the average copy number across all copy-number aberrations called within the annotated gene boundaries based on syn8042988.

The somatic eQTL, ASE and mutational signature analyses also accounted for total somatic mutation burden (number of SNVs and short insertions and deletions (indels)) and sample purity (Supplementary Table 1). Purity was estimated based on copy-number segmentation. In addition, the somatic eQTL and ASE analyses accounted for local SNV burden calculated in a 1-Mb window from the gene coordinates (<https://dcc.icgc.org/api/v1/download?fn=/PCAWG/transcriptome/eQTL/covariates/pergene.somatic.snv.cis.burden.1188.wl.donors.tsv.gz>).

The germline eQTL analysis also modelled the population structure as random effect. The population structure was assessed by a kinship matrix that was calculated based on every twentieth germline variant, processed as described below (see 'Germline eQTL variants'). The kinship matrix was then calculated as an empirical patient-by-patient covariance matrix.

Different covariates were accounted for per-analysis method (Supplementary Table 1). The project code describes cancer type and country-of-origin. Somatic burden is the total number of SNVs and indels. Purity was estimated based on copy-number segmentation. Local somatic burden is the number of SNVs in a 1-Mb window around the gene coordinates. Local copy number was defined as the average copy-number state across all SCNAs called within the annotated gene boundaries.

## GO and Reactome pathway enrichment

We performed GO<sup>74,75</sup> and Reactome pathway<sup>20,21</sup> enrichment with the Bioconductor packages biomaRt<sup>76,77</sup>, clusterProfiler<sup>78</sup> and ReactomePA<sup>79</sup>

(FDR  $\leq 10\%$ ). The number of genes used as background set is described per analysis method.

## Germline eQTL variants

PCAWG variant calls v.0.1<sup>3</sup> were downloaded from GNOS and processed following the PCAWG-8 protocol: (1) VCF files were indexed and merged using bcftools<sup>80</sup>. (2) All variants were filtered for 'PASS' flag. (3) All variants were filtered for quality larger than 20. (4) Only bi-allelic sites were considered.

HDF5 files for each 100-kb chunk of the VCF files were generated, assuming additivity that was numerically encoded as 0, 1 or 2 for homozygous reference, heterozygous or homozygous alternative state, respectively. For indels, we encoded the presence or absence of the variant as 0 or 1, respectively. Each variant was normalized to mean 0 and standard deviation 1. Missing variants were mean-imputed. To create our eQTL release set v.1.0, the resulting HDF5 files were subsequently merged into a global HDF5 file and all variants which follow any of the following conditions were removed: (1) minor allele frequency  $\leq 1\%$ ; and (2) missing values  $\geq 5\%$

## Germline eQTL analysis

In the germline eQTL analyses, we used the processed gene expression dataset from 1,178 patients for which germline variant calls (eQTL release set v.1.0, see 'Germline eQTL variants') were available. Linear mixed models were used to model the correlation between germline variants (within 100 kb of gene boundaries) and gene expression values (see 'Gene expression filtering') using the limix package<sup>81</sup>. Known covariates were modelled as fixed effects and population structure as random effect (see 'Covariates').

A two-step approach was used to adjust for multiple testing. First, for each gene, we adjusted for the number of independent tests estimated based on local linkage disequilibrium<sup>82</sup>. Second, we performed a global correction across the lead variants, that is, the most significant SNPs, per eQTL. Germline eGenes were defined as genes with an eQTL with global FDR  $\leq 5\%$ .

## GTEx comparative analysis

The GTEx comparative eQTL analysis was based on the eQTL maps v.6p<sup>10</sup>. We mapped the positions and alleles of our PCAWG-specific eQTL to the eQTL in all GTEx tissues. To determine whether a lead eQTL variant is replicated in a given GTEx tissue, we followed the previously described strategy<sup>10</sup>. For each eGene, we considered the eQTL lead variant and assessed the replicability of the signal in the GTEx cohort based on marginal association statistics using 42 GTEx tissues without cell lines ( $P < 0.00024 = 0.01/42$ , corrected for the number of GTEx tissues—that is, 42). If the lead variant did not replicate or was not tested, we determined replication based on the variant with the smallest  $P$  value within the linkage disequilibrium block ( $r^2 \geq 0.8$  estimated based on UK10K project) of the lead variant across 25 (or 42) tissue-matched GTEx analyses. If neither lead nor any variant within the linkage disequilibrium block was tested, we determined replication based on the smallest  $P$  value of any variant within the 100-kb window tested within the GTEx cohort. We also derived less stringent sets of PCAWG-specific eGenes by allowing replication in up to 1, 5 or 10 GTEx tissues.

## Tissue sharing of germline eGenes between histotypes

Using the R package qvalue (<https://github.com/StoreyLab/qvalue>, v.2.14.0), we generated  $\pi_1$  statistics comparing the lead variants of one histotype against their  $P$  value distribution in the other histotypes. Because  $\pi_1$  statistics are known to be confounded by sample size and number of eQTL found, we subsampled the eQTL lead variants to a randomly selected set of 100 variants. After 20 rounds of subsampling, we derived the same  $\pi_1$  statistics as mentioned earlier and reported the average.

## Roadmap enrichment of germline eGenes

For each lead variant, we generated a matching background set of 1,000 variants using SNPsnap<sup>83</sup>. Each variant (background and foreground) was intersected with the location of 25 Roadmap factors<sup>16</sup> in 127 cell types. From this we derived fold change and *P* values. Significant changes of fold change between PCAWG-specific and non-specific eQTLs is based on a one-sided Wilcoxon rank-sum test.

## Enrichment analysis

Enrichment of Reactome pathways of PCAWG-specific eGenes was performed using the Bioconductor package ReactomePA<sup>79</sup>.

## Somatic calls and mutational burden

We used the set of consensus SNVs somatic calls provided by PCAWG (syn7357330) based on three core caller pipelines and MuSE<sup>84</sup>. On average, we counted 22,144 somatic SNVs per patient, with different median numbers of SNVs per cancer type, ranging from 1,139 in thyroid adenocarcinoma to 72,804 SNVs in skin melanoma (Extended Data Fig. 5a). Owing to the low frequency of somatic SNVs across the cohort (Extended Data Fig. 5b), we collapsed the variants by genomic regions defined by gene annotations (Gencode v.19<sup>60</sup>). Specifically, we generated a set of disjoint gene exons by collapsing overlapping exon annotations into single features using bedtools<sup>85</sup>. The set of disjoint introns was generated using bedtools by subtracting the collapsed exonic regions from the gene regions. To map local effects of somatic mutations in flanking features outside the gene body, we binned the surrounding regions (plus and minus 1 Mb from the gene boundaries) into 2-kb windows (flanking) overlapping by 1 kb.

We defined three different types of aggregated somatic burden to assess differences in power in detecting somatic eGenes and *P* value calibration. The burden in a genomic region was defined as (1) a binary value that indicates presence or absence of SNVs; (2) the aggregated burden as sum of SNVs; or as (3) weighted burden, that is, sum of variant allele frequencies of the SNVs (Supplementary Fig. 10a) to take into account their clonality (<https://dcc.icgc.org/releases/PCAWG/transcriptome/eQTL/genotypes>). We assessed calibration of all three analyses with Q-Q plots of nominal and permuted *P* values (permutation of the patients in the gene expression matrix) (Supplementary Fig. 10b–d). Moreover, for the linear regression analysis, genotypes were standardized across patients (to mean zero and standard deviation one) and standardized effect sizes are provided in Supplementary Table 5.

Overall, somatic burden within flanking regions was the most prevalent type of burden tested per gene (Extended Data Fig. 6a). We found similar average relative mutation density per type of genomic region (flanking = 0.008 mutations per kb; introns = 0.007 mutations per kb; exons = 0.006 mutations per kb) (Extended Data Fig. 6b) and average recurrence of the same mutated region across the cohort was rather low (flanking = 1.4%; exons = 1.7%; introns = 4%) (Extended Data Fig. 6c).

## Somatic eQTL analysis

Linear models were used to model the correlation between recurrent somatic burden and gene expression of up to 18,898 protein-coding genes, using the limix package<sup>81</sup> (see ‘Gene expression filtering’). Gene expression was corrected for 35 hidden Peer factors. Known covariates were modelled as fixed effects (see ‘Covariates’). We considered only somatic burdens with frequency greater than 1%, including exonic and intronic burdens, as well as flanking burdens, within 1 Mb from gene boundaries.

The somatic eQTL analysis was performed on all 1,188 patients and on the subset of 899 patients with carcinoma (representing 20 of the 27 types of cancer) to replicate the analysis on a more homogeneous set of tumours. A *cis* window of 1 Mb from the gene boundaries was used to find mutated genomic intervals with a burden frequency  $\geq 1\%$  in the cohort (at least 12 patients in the full cohort and 9 patients in

the carcinoma cohort). Together, 18,708 of the genes had at least one mutated interval at that frequency and were included in the analysis and 1,049,102 regions showed a burden frequency  $\geq 1\%$

Benferroni correction was applied to correct for multiple *cis* windows tested within the same gene. Then, Benjamini–Hochberg correction was applied to adjust the *P* values of the lead genomic regions across genes. Somatic eGenes were defined as genes with an eQTL at a FDR  $\leq 5\%$ .

## Somatic cis-eQTL comparative analysis

We compared our 649 somatic eQTL set with three previous cancer studies<sup>86–88</sup> to identify independent evidence of interaction between our eGenes and the associated *cis*-genomic regions with somatic burden. Studies were chosen if they provided lists of cancer regulatory elements linked to genes or regulatory elements with somatic mutations linked to gene expression deregulation in cancer. All the three studies examined were based on TCGA cancers. For this, we checked perfect overlaps with both the somatic burden location and the eGene. Moreover, we looked at the overlap between somatic eQTL and 72,987 GeneHancer<sup>89</sup> enhancers-to-genes interactions, with at least two independent supporting methods (called ‘double-elite’), downloaded from the UCSC hg19 GeneHancer track<sup>90</sup>. We then compared this overlap with a set of nulls generated by 1,000 random permutations of the GeneHancer regulatory elements with nearby genes located within 1 Mb. We then retrieved an empirical *P* value of enrichment by counting the number of random nulls (*N*) showing greater number of overlaps than those found between the somatic eQTL set and the GeneHancer set ( $P = (N + 1) / (1,000 + 1)$ ).

## Functional enrichment in somatic cis-eQTL

To identify putative regulatory sites enriched for somatic eQTL, we retrieved functional annotations of the lead genomic flanking intervals of the somatic eQTL (556 intervals linked to 638 somatic eQTL). Therefore, we mapped somatic eQTL to 25 Roadmap Epigenomics chromatin marks of 127 different cell types<sup>16</sup> and ENCODE transcription-factor binding site annotations in 9 cell types (including 8 cancer and one embryonic stem-cell lines<sup>91</sup>) (Supplementary Tables 6 and 7). We compared annotations in the significant set of eQTLs with a null distribution based on 1,000 random samplings of a matched set of genomic intervals. To define the matched sets of genomic intervals, we selected flanking genomic intervals from the whole set of tested genes that showed a similar distance from the gene start (exact distance  $\pm 2$  kb) and that matched the exact burden frequency of the corresponding interval in the significant associations. We then overlapped the 1,000 matched sets with Roadmap Epigenomics and ENCODE annotations. To avoid ambiguous overlaps (with multiple annotations), we retained only genomic intervals showing a minimum overlap of 10% of their length.

We retrieved an empirical *P* value of enrichment for each annotation by counting the number of randomly sampled flanking intervals (*N*) showing greater number of overlaps compared to the eQTL set ( $P = (N + 1) / (1,000 + 1)$ ). Benjamini–Hochberg correction was applied to the empirical *P* values (over 25 marks in 127 cell lines for Roadmap Epigenomics annotations and over 149 transcription-factor-binding sites for 9 ENCODE cell lines). We then computed the fold change per annotation and cell line as a ratio of annotated lead flanking intervals and mean number of annotated matched random flanking intervals over the 1,000 samplings.

Furthermore, we performed GO<sup>74,75</sup> and Reactome pathway<sup>20,21</sup> enrichment with the Bioconductor packages biomaRt<sup>76,77</sup>, clusterProfiler<sup>78</sup> and ReactomePA<sup>79</sup> (FDR  $\leq 10\%$ ) and also looked at enrichment within high-confidence cancer testis genes previously described<sup>92</sup>, using 18,708 genes with at least one mutated interval as background.

## Variance component analysis

Limix was used to perform variance decomposition using the same covariates as in the somatic variant analyses except for local



copy-number state (see ‘Covariates’). The random effects were based on the following common germline variants and somatic burden (frequency > 1%) (see ‘Somatic calls and mutational burden’ for detailed description of burden): (1) *cis*-somatic intronic: weighted burden in introns; (2) *cis*-somatic exonic: weighted burden in exons; (3) *cis*-somatic flanking: weighted burden in 1-kb-overlapping regions of 2 kb within 1 Mb from gene boundaries; (4) somatic intergenic: weighted burden in 1-kb-overlapping regions of 2 kb outside the 1 Mb window; (5) *cis*-germline: germline variants within 100 kb from gene boundaries; (6) *trans*-germline: genome-wide population structure (see ‘Covariates’); and (7) local copy-number variation (see ‘Covariates’).

All the data was mean-centred and standardized. For each of the random effects, a linear kernel was computed and used as covariance matrix. The resulting variance components were normalized to add up to one.

## Mutational signature associations

We obtained 39 mutational signatures from PCAWG-7 beta 2 release<sup>9</sup> and used linear models to associate the mutational signatures with gene expression of up to 18,898 protein-coding genes across 1,159 patients while accounting for known covariates (see ‘Covariates’) (quality control) (Extended Data Fig. 10a–e). The 1,159 patients were a subset of the total 1,188 patients, for whom mutational signature profiles were available. Gene expression was corrected for 35 hidden peer factors (see ‘Gene expression filtering’).

We retained 18,888 genes that showed a minimum FPKM of 0.1 in at least 1% of 1,159 the patients (see ‘Gene expression filtering’). Signatures with zero variance and a prevalence below 1% were filtered, and we obtained 28 signatures. We applied linear models to associate expression of these genes with the signatures across all 1,159 patients, a subset of 877 patients with carcinoma or a subset of 891 European patients to assess consistency of the associations (Extended Data Fig. 10f, g).

Across all patients, we found 1,176 significantly associated genes after Benjamini–Hochberg correction (we used an FDR ≤ 10% for enrichment analyses, multiple testing was applied across all signature–gene pairs) (Supplementary Tables 19a–c). We performed gene enrichment analyses of the significant genes per signature (see ‘GO and Reactome pathway enrichment’) (here 18,831 background genes, multiple testing correction across all ontologies per signature FDR ≤ 10%) (Supplementary Table 19d). Whereas most signatures were associated with only few genes, 18 showed recurrent *trans* effects and affected expression of over 20 genes (Extended Data Fig. 11d, Supplementary Table 19e). We further found that the vast majority of genes (85.8%) were associated with only one signature (1,009 genes); 129 genes were associated with two, 32 with three, 5 with four and 1 with five signatures.

To assess how tissue-specific both mutational signatures and their associations with gene expression are, we analysed the occurrence of each signature in each of the types of cancer. We assessed the presence (at least one SNV of a signature in at least one patient with a specific cancer type) and mean prevalence (mean number of SNVs of a certain signature across all patients of a specific cancer type) of the signatures in the types of cancer (Extended Data Fig. 13c, d). We defined cancer-type-specific signatures to occur in up to four types of cancer (signatures 4, 7, 9, 12, 16, 38 and 39) and common signatures to be missing in up to five types of cancer (signatures 2, 13 and 18). For each of these signatures, we performed cancer-type-specific analyses, that is, we assessed the association between the respective signature and gene expression in just the patients who are of a cancer type that shows mutations of the respective signature (Extended Data Fig. 13c, left heat map). We then correlated the *P* values of these cancer-type-specific analyses with the *P* values of the analysis across all patients and calculated the Pearson correlation coefficients (Supplementary Fig. 24a–e). We show that the correlation between cancer-type-specific and whole-cohort *P* values is dependent on the sample size of the respective analysis ( $r^2 = 0.671$ ) (Supplementary Fig. 1f).

We further performed PCA on the signatures across both, patients (PCA on signature-specific SNVs per patient) and genes (PCA on adjusted *P* values of signature–gene expression associations) (Extended Data Fig. 11a, b).

To assess significance of the functional annotation of SNVs by mutational signatures, we also associated gene expression with the total number of SNVs and correlated the *P* values ( $-\log_{10}(P)$ ) of the associations with the respective signature-specific *P* values. The absolute Pearson correlation coefficients remain below 0.1 (Supplementary Table 19f).

To establish causality of signature–gene expression associations, we included the germline eQTL into the analysis using linear mixed models; 197 of our 1,176 signature-associated genes were also germline eGenes. These 197 associations involved 26 of the 28 mutational signatures. We associated the lead variants of these eGenes with the rank-standardized signature SNVs across 2,507 patients. We used the subset of the 2,818 WGS patients for which mutational signature profiles and all known covariates were available. We accounted for the same fixed covariates as in the mutational signature–gene expression association studies and, in addition, for kinship as a random effect (see ‘Covariates’).

We then performed proportional colocalization analysis with Bayesian model averaging using the R package coloc<sup>93</sup> to test whether gene expression and mutational signatures share common causal genetic variants in a given gene region. A proportional colocalization analysis tests the null hypothesis of colocalization by assuming that two phenotypes that share causal variants will have proportional regression coefficients for either phenotype with any variant selection in the vicinity of the causal variant. We applied the Bayesian model averaging approach, with each tested model consisting of a selection of two variants. The *P* values are then averaged over all models to generate posterior predictive *P* values<sup>93</sup>. We filtered variants so that no pair of variants showed  $r^2 > 0.95$  and each variant’s marginal posterior probability of inclusion with one of the phenotypes was greater than 0.01. The nominal *P* values of rejecting the null hypothesis of colocalization are listed in Supplementary Table 19e.

We then performed mediation analysis<sup>94,95</sup> to assess directionality of the effect between germline eQTL, gene expression and mutational signature. First, causal mediation analysis was applied to each of the triples of eQTL lead variant, gene and mutational signature using a structural equation model from the R package lavaan<sup>96</sup>. Then, we used the R package mediation<sup>97</sup> to assess significance of mediation and estimate the proportion of mediated effect by non-parametric bootstrap confidence intervals (1,000 simulations).

## ASE analysis: assembling phased germline and somatic variants

To understand the precise effect of somatic variations in their genomic context and for subsequent allele-specific analyses, both germline and somatic variants were phased. For assembling phased germline genotypes, we used the Sanger 1000G callset<sup>6</sup>, and applied IMPUTE2<sup>98</sup> for phasing of heterozygous germline variants. The IMPUTE2 output was corrected using results from the Battenberg CN calling algorithm<sup>99</sup> to ascertain that no haplotype switches occur within regions of consecutive copy-number gain. The resulting phased germline genotypes were arranged such that haplotype 1 always corresponded to the amplified alleles in regions with SCNAs (major allele). In cases in which both co-occur on the same NGS read (approximately 10 million variants, 20% of all SNVs), we phased individual somatic variants to the nearest germline heterozygous site. For downstream analyses, we considered only SNVs that were phased by at least three reads to the respective germline variant (approximately 6 million out of 10 million SNVs).

All phased SNVs were aggregated into functional categories based on their genomic regions defined by gene annotations (upstream, downstream, promoter, 5′ UTR, intron, synonymous, missense, stop gain and 3′ UTR) and mapped to the nearest gene within a *cis* window of 100 kb using the Variant Effect Predictor (VEP) tool<sup>100</sup>. Promoter

variants were defined as 1-kb upstream of the TSS. We included flanking regions by using the VEP 'UpDownDistance' plugin with a maximum range parameter of 100 kb. We divided the upstream and downstream variant categories into disjoint categories using 10-kb windows from 10 to 100 kb. We integrated 'splice donor' and 'splice acceptor' variants into the general 'splice region' variant category and mapped 'stop retained' variants to the 'synonymous' variant category. We averaged transcript-level annotations to gene-level annotations to retrieve the expected functional effect of a variant for a given gene. We analysed the relationship between SNV variant allele frequency and SCNAs at the same locus to determine whether variants occurred before ('early') or after ('late') the corresponding SCNA (PCAWG-11). We computed a weighted *cis*-mutational burden per category by estimating the cancer cell fraction of each SNV and aggregating SNVs to a total localized burden weighted by their respective cancer cell fraction.

### ASE read counts

The positional information of the heterozygous germline variants was used together with the RNA-seq BAM files as input to the GATK ASEReadCounter<sup>101</sup> algorithms for counting ASE reads. We considered reads with a minimum mapping quality of 20 and a minimum base quality of 10. Only heterozygous variants with a minimum coverage of eight RNA-seq reads were considered for all further analyses.

The raw ASE read counts were post-processed as follows: (1) ASE sites were converted to BED files and aligned against the ENCODE 50-mer mappability track (wgEncodeCrgMapabilityAlign50mer.bigWig) to extract mappability scores for all sites. All sites with mappability scores unequal to 1 were removed. (2) All sites with allelic read counts less or equal to 1 were removed to prevent genotyping error to influence ASE quantification. (3) All sex chromosomes were dropped for further analysis. (4) We estimated sequencing error per patient as the sum of non-reference and non-alternative bases over the total number of bases. We assessed statistical mono-allelicity through a binomial test using the estimated sequencing error probabilities, corrected using the Benjamini–Hochberg step down procedure. All sites that appeared to be statistically mono-allelic were removed. (5) For each ASE site, copy-number states were retrieved from the Sanger copy-number consensus callset (PCAWG-11). Purity estimates for each patients were retrieved from the accompanying purity tables.

To aggregate site-level ASE to a gene-level readout and to allow for estimation of effect directionality, we used the phased germline genotypes. Gene mapping was performed against ENSEMBL release 75 using the pyEnsembl Python library. We retrieved all genes at each ASE site and summed up the read counts on the respective haplotypes to gene-level haplotype-specific read counts. We further averaged haplotype-specific copy-number states to a mean haplotype-specific copy-number state per gene and computed the gene-level copy-number ratio as the major over total ratio of those averages. To allow for a robust assessment of gene-level ASE, we considered only genes with at least 15 reads total, yielding 4,379,378 gene–patient pairs of 1,120 patients and 17,009 unique genes across 12,441,502 accessible sites in total. Every remaining gene was tested for AEI using a binomial test against an expected read ratio of 0.5 to derive nominal *P* values, and a binomial test against the expected copy-number ratio modified by tumour purity to derive copy-number-corrected *P* values. Nominal and copy-number-corrected *P* values were adjusted separately for multiple testing using the Benjamini–Hochberg procedure. Significant AEI was called at FDR ≤ 5%. We further annotated each gene with the number of ASE sites used for aggregation. For all downstream analyses, we considered only genes annotated as protein coding (ENSEMBL biotype = 'protein\_coding').

### Generalized linear models

Across all 4,379,378 gene–patient pairs, we trained multivariate linear models using (i) logistic regression against a binary indicator of AEI absence or presence in a gene, or (ii) standard linear regression against

the phased ASE ratio of a gene to assess the directionality of the regulatory change. For (i), haplotype-specific mutations were summed up to a total burden per category, whereas for (ii) we used the difference in burden between the haplotypes 1 and 2. The consistency of the phasing map between somatic variants and ASE sites ensured that model coefficients kept their directionality independent of the arbitrary labelling of haplotypes as 1 or 2. The full set of considered factors is as follows: (1) copy-number ratio at the gene locus ( $0.5 \leq x \leq 1$ ); (2) sample purity ( $0 < x < 1$ ); (3) natural logarithm of total gene length ( $x > 0$ ); (4) natural logarithm of the length of the canonical transcript ( $x > 0$ ); (5) heterozygosity of the lead eQTL variant ( $x = 0$  if homozygous,  $x = 1$  if not homozygous); (6) all mutational burden categories as determined by VEP annotations (upstream in 10-kb windows, downstream in 10-kb windows, promoter, 5' UTR, intron, synonymous, missense, stop gain and 3' UTR;  $x \geq 0$  for logistic model,  $x \in \mathbb{R}$  for directed model).

To compare global effects and different contributions of SCNA, germline eQTL, coding and non-coding SNVs, a simplified logistic model was trained after accumulating all coding and non-coding variants to separate categories and reporting standardised effect sizes (Fig. 1e).

### Cancer gene enrichment

Cancer gene enrichment was conducted on the COSMIC census<sup>53</sup> using Fisher's exact test and gene set enrichment analysis as previously described<sup>102</sup>. For enrichment, the average score of a gene was computed across the cohort and only genes with at least five replicates in the cohort were kept, yielding a total of 16,078 genes.

### Chromosomal distribution of ASE

We calculated the recurrence of ASE genes in each tumour type. To examine the chromosomal distribution of ASE genes, we calculated the average recurrence of all genes for every 200-gene window with a 10-gene step, and then subtracted the average ASE occurrence in each tumour type to obtain the peaks of ASE surplus across all chromosomes. The recurrence of copy-number genes was calculated in an analogous manner.

### Estimation of alternative promoter activity

We estimated promoter activities using RNA-seq data and Gencode (release 19) annotations for 70,937 promoters in 20,738 genes. We grouped transcripts with overlapping first exons under the assumption that they are regulated by the same promoter<sup>103</sup>. TSSs that are located within internal exons, or which overlap with splice acceptor sites, were removed from this analysis as these promoters are difficult to estimate from RNA-seq data<sup>28</sup>. Promoter activity can be estimated using exon usage<sup>29</sup>, spliced reads<sup>28</sup> or isoform-based estimates<sup>30</sup>. Here we used an isoform-based approach to quantify promoter activity. We quantified the expression of each transcript from the RNA-seq data using Kallisto<sup>66</sup> and calculated the sum of expression of the transcripts initiated at each promoter to obtain an estimate of promoter activity. To obtain the relative activity for each promoter, we normalized each promoter's activity by the overall gene's expression. We divided the promoters of each gene into three categories based on their average pan-cancer promoter activity. The promoters with <1 FPKM average activity are called inactive promoters, and the most active promoter of each gene is called the major promoter. The remaining active promoters of the gene are called minor promoters.

The association between promoter activities and promoter mutation burden was estimated using the same framework as the somatic eQTL analysis. We examined associations for the promoters of expressed multi-promoter genes with a burden frequency ≥ 1% in the cohort (at least 12 patients in the full cohort). The weighted burden of the region 1-kb upstream of the TSS—that is, the sum of variant allele frequencies of the SNVs for each gene—was used as the genotype for the promoters of the respective genes. We used linear models to study the associations between the recurrent somatic burden and the promoter activity (both

# Article

for the relative activity and the  $\log_2$ -transformed absolute activity). Similar to the somatic eQTL analysis, the known covariates and the 35 hidden peer factors were provided as cofactors to the linear models. We adjusted the *P* values using Benjamini–Hochberg correction method and looked for associations with  $\text{FDR} \leq 5\%$ .

## Identification of alternative splicing

We used the alignments based on the STAR pipeline to collect and quantify alternative splicing events with SplAdder<sup>70</sup>. The software has been run with its default parameters with confidence level 3. We generated individual splicing graphs for each RNA-seq sample for both tumour samples as well as matched healthy samples (when available). All graphs were then integrated into a merged graph to comprehensively reflect all splice junctions observed in all samples together. On the basis of this combined graph, SplAdder was used to extract alternative splicing events of the following types: alternative 3' splice site, alternative 5' splice site, cassette exon, intron retention, mutually exclusive exons, coordinated exon skip (see supplementary figure 3 in ref. <sup>70</sup>). Each identified event was then quantified in all samples by counting split alignments for each splice junction in any previously identified event and the average read coverage of each exonic segment involved in the event was determined. We then computed a PSI value for each event that was then used for further analysis. We further generated different subsets of events, filtered at different levels of confidence, in which confidence is defined by the SplAdder confidence level (generally 2), the number of aligned reads supporting each event, the number of samples that were found to support the event by SplAdder, and the number of samples that passed the minimum aligned read threshold.

## Enrichment of outlier splicing associated with splice sites and branchpoint motifs

We assessed the significance of mutational enrichment for 5' and 3' splice sites, and branch-point<sup>104,105</sup> intronic regions using a permutation-based approach. Impactful mutations were defined as mutations overlapping exons and introns involved in cassette exon events, in which the PSI-derived *z*-score was  $\geq 3$  or  $\leq -3$ . For each intronic site, we compared the frequency of observed impactful mutations against frequencies of randomly sampled intronic regions (number of iterations = 1,000). For exonic sites, the null distribution was established from randomly sampled exonic sites. Randomly sampled sites were within a 100-bp window around the 5' and 3' splice site. For branch-point regions, sampled sites were within a 50-bp window around the branch-point sequence. The *P* value was computed as the number of randomly sampled frequencies greater or equal to the observed frequency.

## SAVNet analysis for identifying rare SAVs

The SAVNet approach<sup>35</sup> was designed for identifying somatic variants associated with local aberrant splicing alterations from matched genome and transcriptome sequencing data. It uses permutations to calculate an FDR and by restricting to two classes of relationships between somatic mutations and splicing alterations to focus: (1) splice site disruption, in which exon skipping, alternative 5' or 3' splice site, or intron retention is associated with a mutation in a splice site motif; and (2) splice site creation, in which alternative 5' or 3' splice sites are associated with mutations that create a novel splice motif ( $\text{FDR} \leq 10\%$ ) (Extended Data Fig. 17e).

## Identification of RNA fusions

Gene fusions between any two genes were identified based on two gene fusions detection pipelines: FusionMap (v.2015-03-31) pipeline<sup>106</sup> and FusionCatcher (v.0.99.6a)/STAR-Fusion (v.0.8.0) pipeline<sup>107</sup>. ChimerDB 3.0 was used as a reference of previously reported gene fusions. The database contains 32,949 fusion genes split into three groups: (1) KB: 1,067 fusion genes manually curated based on public resources of fusion genes with experimental evidences; (2) Pub: 2,770 fusion genes

obtained from text mining of PubMed abstracts; and (3) Seq: archive with 30,001 fusion gene candidates from deep-sequencing data. This set includes fusions found by re-analysing the RNA-seq data of the TCGA project encompassing 4,569 patients from 23 types of cancer.

In brief, FusionMap was applied to all unaligned reads from the PCAWG aligned TopHat2 RNA-seq BAM files for each aliquot to detect gene fusions. In the FusionCatcher/STAR-Fusion pipeline, for each aliquot with paired-end RNA-seq reads FusionCatcher was applied to the raw reads, with the genome reference. Specifically, for each aliquot with paired-end RNA-seq reads FusionCatcher was applied to the raw reads. The '-U True; -V True' runtime options were used. For each aliquot with single-end RNA-seq reads, STAR-Fusion was applied to the raw reads, with the same reference genome and gene models as FusionCatcher and with default settings. In parallel, FusionMap was applied to all unaligned reads from the PCAWG aligned TopHat2 RNA-seq BAM files for each aliquot to detect gene fusions with the following non-default options values: MinimalHit = 4; OutputFusionReads = True; RnaMode = True; FileFormat = BAM.

To reduce the number of false-positive fusions, the two sets of fusions were filtered to exclude fusions based on the number of supporting junction reads, sequence homology, and occurrence in normal samples (from the GTEx and the PCAWG cohort). To get a high-confident consensus fusion call set from these two pipelines, a fusion to be included in the final set of fusions had to: (i) be detected by both fusion detection tools in at least one sample; and/or (ii) be detected by one of the methods and have a matched structural variant in at least one sample. The consensus WGS-based somatic structural variants (v.1.6) were obtained from the PCAWG repository in <https://dcc.icgc.org/releases/PCAWG>.

For integration with matched structural variant evidence, a fusion was considered to match a structural variant if the absolute distance between the fusion break points and structural variant break points did not exceed 500 kb (the distance was considered infinite when the chromosomes of the fusion and structural variant break point differ). When there was no evidence for a direct structural variant fusion, the search was expanded to look for composite fusions. In this case, an exhaustive search was performed to look for two structural variants with break points close to the fusion break points and with an effective distance smaller than 250 kb.

Finally, 3,540 fusion events were included as the consensus fusion call set, from these 2,268 were detected by both FusionCatcher/STAR-Fusion and FusionMap (from these, 1,821 had matched structural variant evidence) and 1,112 were detected by only one method and had matched structural variance evidence.

In total, approximately 36% of all detected fusion transcripts were predicted to be in-frame, several UTR-mediated fusion transcripts preserve complete coding sequences of one fusion partner. These include a known fusion *TBL1XR1-PIK3CA* in a breast tumour and a notable new example *CTBP2-CTNBN1* in a gastric tumour.

All fusions are available in Synapse: <https://dcc.icgc.org/releases/PCAWG/transcriptome/fusion>.

## Identification of RNA-editing events

We used an RNA-editing events calling pipeline, which is an improved version of that previously published<sup>108</sup>. First, we summarized the base calls of pre-processed aligned RNA reads to the human reference in pileup format. Second, the initially identified editing sites were then filtered by the following quality-aware steps: (1) the depth of candidate editing site, base quality, mapping quality and the frequency of variation were taken into account to do a basic filter: the candidate variant sites should be with base-quality  $\geq 20$ , mapping quality  $\geq 50$ , mapped reads  $\geq 4$ , variant-supporting reads  $\geq 3$ , and mismatch frequencies (variant-supporting-reads/mapped-reads)  $\geq 0.1$ . (2) Statistical tests based on the binomial distribution  $B(n, p)$  were used to distinguish true variants from sequencing errors on every mismatch site<sup>109</sup>, in which *p* denotes the background mismatch rate of each transcriptome

sequencing, and  $n$  denotes sequencing depth on this site. (3) Discard the sites present in combined DNA SNP datasets (dbSNP v.138, 1000 Genome SNP phase 3, human Dutch populations<sup>110</sup>, and BGI in-house data; combined datasets deposited at: <ftp://ftp.genomics.org.cn/pub/icgc-pcawg3>). (4) Estimate strand bias and filter out variants with strand bias based on two-tailed Fisher's exact test. (5) Estimate and filter out variants with position bias, such as sites only found at the 3' end or at 5' end of a read. (6) Discard the variation site in simple repeat region or homopolymer region or <5 bp from splicing site. (7) To reduce false positives introduced by misalignment of reads to highly similar regions of the reference genome, we performed a realignment filtering. Specifically, we extracted variant-supporting reads on candidate variant sites and realign them against a combination reference (hg19 genome plus Ensembl transcript reference v.75) by bwa0.5.9-r16. We retain a candidate variant site if at least 90% of its variant-supporting reads are realigned to this site. Finally, all high confident RNA-editing sites were annotated by ANNOVAR<sup>111</sup>. (8) To remove the possibility of an RNA-editing variant being a somatic variant, the variant sites are positionally filtered against PCAWG WGS somatic variant calls (9). The final two steps of filtering are designed to enrich the number of functional RNA editing sites. First, we keep only events that occur more than two times in at least one cancer type. Second, we keep only events that occur in exonic regions with a predicted function of missense, nonsense or stop-loss. The final step of filtering within exonic regions with a specific predicted function induces the largest difference in observed frequencies of RNA-editing events between our analysis and the published one<sup>108</sup>. A comparative depiction of the frequencies of RNA-editing events identified in our analysis (Supplementary Table 24) and the previously published analysis<sup>108</sup> is seen in Supplementary Fig. 23.

### Gene-centric table creation

To perform joint analysis across RNA and DNA alterations, each alteration type was condensed into a binary gene-centric format. Because alterations occur at many different scales (nucleotide, exonic, gene or transcript), to make them comparable we projected each alteration type onto the gene body. We summarized each alteration type by its presence or absence within a single gene, yielding a binary value per type for each gene-sample pair.

The events we included in this analysis were: RNA editing, non-synonymous variants, expression, splicing alterations, copy-number alterations, fusions and alternative promoters. Each alteration type was summarized differently owing to their inherent differences.

RNA-editing events and non-synonymous variants can occur several times within a single gene body, so these events were denoted as 1 if they occurred at least once within a gene-sample pair.

For copy number, to obtain a single numerical value per gene-sample pair, the copy-number alteration was averaged over the gene body. Because we do not have matched normal samples against which to compare, we instead consider outlying events within each histotype as significant. Thus, a value of 1 was given to average copy-number alterations larger than 6 or smaller than 1.

Similar to non-synonymous variants, multiple splice events can occur within a gene body. The event with the most extreme PSI value within the gene body is selected as the candidate event for the gene. The candidate's PSI value for a gene is compared over all samples within a histotype and it is set to 1 (that is, significant) only if the absolute value of its z-score is larger than 6 and the standard deviation is larger than 0.01 within that histotype.

Similar to expression outliers, we calculate a z-score using the log-transformed upper-quartile normalized FPKM values with a pseudo-count of 1. All genes within a histotype with a standard deviation larger than zero and an absolute value larger than three were identified as an outlier. Alternative promoter outliers were calculated based on relative promoter activity within each cancer type. To binarize the promoter

activity, a z-score cut-off of two over the relative expression distribution within each cancer type was used.

For ASE outliers, only genes with significant allelic imbalance (FDR  $\leq 5\%$  and allelic imbalance  $> 0.2$ , binomial test) were denoted as 1. All ASE events that were identified were further filtered to keep only genes that have not been identified as imprinted<sup>26</sup>.

In addition to the z-score-filtering mentioned above, we further filtered non-synonymous SNVs, RNA-editing events and splicing events such that they either induce a frameshift or the alternative region contains an HGMD variant<sup>112</sup> of the category 'damaging'.

It must be noted that in many cases, the z-score calculated is not from a Gaussian distribution, so some events may be missed or falsely included. Through our choice of very stringent z-score thresholds and functional filters, we hope that spurious outlier events are minimized.

### Pathway analysis

For our pathway analysis, we used the TCGA pathway definitions to examine genes and pathways that have several alterations at both the DNA and RNA level<sup>113</sup>.

### Co-occurrence analysis

The co-occurrence analysis was also performed on the aforementioned binarized gene-centric table, but only including variants, expression outliers, alternative promoters, alternative splicing and fusions. SCNA and ASE are excluded owing to a large number of anticipated co-occurrence. In this analysis, we required at least one gene of a given alteration pair to be a COSMIC gene. For each alteration pair, based on the number of donors with both alterations, one alteration only and neither alterations in a set of cancer samples, we performed Fisher's exact test to determine whether the alteration pair was independent of each other. Such tests were followed by Benjamini-Hochberg multiple testing correction to obtain the FDR (or  $q$  values). To rule out the potential false-positive association caused by tissue-specific alterations, we performed the same analysis for each of the tumour types with at least 50 patients, and retained only those alteration pairs that were significantly associated in both the pan-cancer analysis and in at least one specific cancer indication. Among the significantly associated alteration pairs, the co-occurred pairs were those with odds ratio greater than 1. Pathway enrichment and visualization<sup>21,114</sup> were conducted using the R package ReactomePA<sup>79</sup>. The circos plots were generated using the R package circlize<sup>115</sup>. The splicing related genes were derived from the genes annotated as 'REACTOME\_MRNA\_SPLICING' or 'REACTOME\_MRNA\_SPLICING\_MINOR\_PATHWAY' in the Molecular Signatures Database (MSigDB)<sup>116</sup>.

### Identifying genes with heterogeneous mechanisms of alterations in *cis*

Genes with multiple heterogeneous mechanisms of RNA alteration were identified from associations of *cis* variants with gene expression, ASE, fusions and splicing. For gene expression, genes associated with somatic eQTL with FDR  $< 5\%$  were selected. For ASE, the top 5% of genes ranked by the predicted contribution of somatic variants on ASE. For fusions, all RNA fusions with structural variant support were selected. For splicing, genes having somatic mutations within 10 bp of an annotated splice site or 3 bp of a branch point and associated splicing were selected. These associated splicing events also had to have a |z-score| greater than or equal to 3 and the difference of percent spliced in the outlier event was greater than or equal to 10%.

### Recurrence analysis

The recurrence analysis was performed on the binarized gene-centric table for all nine alteration types. The recurrence analysis was performed in three main steps: (1) Aggregate within each alteration type across all samples. This results in a sum for each gene-alteration pair. (2) Convert the counts to ranks within each alteration. The smallest rank

# Article

goes to the most frequently altered genes. Ranks are split evenly across ties. (3) To generate a single score for each gene, the second smallest rank across alterations is used as the score. To identify a score cut-off value for significantly altered genes, a null distribution was generated through permutation. The permutations were performed over the samples within each gene-alteration pair, this was done over all genes and samples 1,000 times, concatenating together all observations, results in 16.8 million permuted scores.  $P < 0.05$  as derived from the null distribution was defined as significant, resulting in a score greater than or equal to 774 considered as significant.

WEX<sup>T17</sup> was used to test the significance of mutually exclusivity of RNA and DNA alterations. As further evidence that *CDK12* alterations may have a functional affect, we find evidence of the previously detected link<sup>55</sup> between a large tandem duplicator phenotype (here defined as more than 10 tandem duplications of size greater than 100 kb) and *CDK12* somatic eQTL mutation (7 out of 18 somatic eQTL carriers are also among the 215 large tandem duplicator cases,  $P = 0.032$ , hypergeometric test).

## Statistical tests

All common statistical tests are two-sided unless otherwise specified. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, and other core data generated by the ICGC and TCGA PCAWG Consortium are described in an accompanying Article<sup>5</sup> and are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA data access committee via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC data access compliance office (<http://icgc.org/daco>) for the ICGC portion of the dataset. In addition, to access somatic SNVs derived from TCGA donors, researchers will also need to obtain dbGaP authorization. Data derived specifically from RNA-seq analysis can be found at <https://dcc.icgc.org/releases/PCAWG/transcriptome>. Subfolders contain identification and quantification of alternative promoter usage, alternative splicing, RNA fusions, gene expression, transcript-level expression and RNA editing. Identified eQTLs are in <https://dcc.icgc.org/releases/PCAWG/transcriptome/eQTL> and a binarized table indicating all RNA and DNA alterations for each gene can be found in the subfolder [https://dcc.icgc.org/releases/PCAWG/transcriptome/recurrence\\_analyses/](https://dcc.icgc.org/releases/PCAWG/transcriptome/recurrence_analyses/). In addition, quality-control metrics and metadata are also included. Some datasets are denoted with synXXXXX accession numbers and available at Synapse (<https://www.synapse.org/>).

## Code availability

The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at <https://dockstore.org/search?search=pcawg> under the GNU General Public License v.3.0, which allows for reuse and distribution. Further details on code availability are in the Supplementary Information.

58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Kim, D. et al. TopHat2: accurate alignment of transcriptsomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
60. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
61. Fonseca, N. A., Petryszak, R., Marioni, J. & Brazma, A. iRAP - an integrated RNA-seq analysis pipeline. Preprint at <https://www.biorxiv.org/content/10.1101/005991v1> (2014).
62. Bioinformatics, B. FastQC: a quality control tool for high throughput sequence data; <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2011).
63. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
64. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
66. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
67. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
68. Krijthe, J. H. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation; <https://github.com/jkrijthe/Rtsne> (2015).
69. Dettro, S. C. et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. Preprint at <https://www.biorxiv.org/content/10.1101/312041v4> (2018).
70. Kahles, A., Ong, C. S., Zhong, Y. & Ratsch, G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**, 1840–1847 (2016).
71. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
72. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protocols* **7**, 500–507 (2012).
73. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
74. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
75. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
76. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protocols* **4**, 1184–1191 (2009).
77. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
78. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
79. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
80. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
81. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. Preprint at <https://www.biorxiv.org/content/10.1101/003905v2> (2014).
82. Davis, J. R. et al. An efficient multiple-testing adjustment for eQTL studies that accounts for linkage disequilibrium between variants. *Am. J. Hum. Genet.* **98**, 216–224 (2016).
83. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2015).
84. Fan, Y. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
85. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
86. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, 362 (2018).
87. Zhang, W. et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).
88. Smith, K. S. et al. Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. *Nucleic Acids Res.* **43**, 5307–5317 (2015).
89. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, 2017 (2017).
90. Haussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
91. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
92. Wang, C. et al. Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types. *Nat. Commun.* **7**, 10499 (2016).
93. Wallace, C. Statistical testing of shared genetic control for potentially related traits. *Genet. Epidemiol.* **37**, 802–813 (2013).
94. Baron, R. M. & Kenny, D. A. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).
95. Preacher, K. J. & Hayes, A. F. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav. Res. Methods Instrum. Comput.* **36**, 717–731 (2004).
96. Rosseel, Y. lavaan: An R Package for structural equation modeling. *J. Stat. Softw.* **48**, 2 (2012).
97. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R Package for causal mediation analysis. *J. Stat. Softw.* **59**, 5 (2014).



98. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
99. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
100. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
101. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
102. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
103. Frith, M. C. et al. A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).
104. Signal, B., Gloss, B. S., Dinger, M. E. & Mercer, T. R. Machine learning annotation of human branchpoints. *Bioinformatics* **34**, 920–927 (2018).
105. Mercer, T. R. et al. Genome-wide discovery of human splicing branchpoints. *Genome Res.* **25**, 290–303 (2015).
106. Ge, H. et al. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27**, 1922–1928 (2011).
107. Nicorici, D. et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Preprint at <https://www.biorxiv.org/content/10.1101/011650v1> (2014).
108. Han, L. et al. The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell* **28**, 515–528 (2015).
109. Li, Q. et al. Caste-specific RNA editomes in the leaf-cutting ant *Acromyrmex echinator*. *Nat. Commun.* **5**, 4943 (2014).
110. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
111. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
112. Stenson, P. D. et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).
113. Sanchez-Vega, F. et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337.e10 (2018).
114. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
115. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
116. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
117. Leiserson, M. D. M., Reyna, M. A. & Raphael, B. J. A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics* **32**, i736–i745 (2016).
118. Rafnar, T. et al. Sequence variants at the *TERT-CLPTM1L* locus associate with many cancer types. *Nat. Genet.* **41**, 221–227 (2009).
119. Bojesen, S. E. et al. Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **45**, 371–384 (2013).
120. Ye, K. et al. Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.* **22**, 97–104 (2016).

**Acknowledgements** Funding for this work was provided by the Damon Runyon Cancer Research Foundation (A.N.B.), European Research Council (RNAEDIT-649019, Q.-P.-H.). C.M.S. was supported by National Institutes of Health (NIH) training grants T32GM008646 and 2R25GM058903. K.-V.L., A.K., N.R.D., S.G.S. and G.R. received core funding from ETH Zurich and MSKCC (New York). This work was also partially supported by SPHN/PHRT Project (106 to G.R.). L.U., R.F.S. and O.S. received support from core funding of the EMBL and the EU

Horizon2020 research and innovation programme (grant agreement N635290). R.F.S. and J.M. received support from the Helmholtz Foundation and the Max Delbrueck Center for Molecular Medicine. Y.H., F.L., F.Z. and Z.Z. received support from Beijing Advanced Innovation Centre for Genomics at Peking University, Key Technologies R&D Program (2016YFC0900100), National Natural Science Foundation of China (81573022, 31530036, 91742203). C.C., L.G., N.F. and A.B. received support from core funding of the EMBL and from EU FP7 Programme projects EurocanPlatform (grant agreement 260791) and CAGEKID (241669). J.G. received support from the Agency for Science, Technology and Research (A\*STAR). D.D. received support from the Singapore International Graduate Award (SINGA) and A\*STAR. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

**Author contributions** The design of the study was contributed by C.C., N.R.D., D.D., N.A.F., Y.H., A.K., K.-V.L., F.L., Y.S., C.M.S., L.U., L.G., S.L., D.L., M.D.P., Q.X., F.Z., J.Z., P.B., S.E., K.A.H., Y.H., M.R.H., H.K., J.O.H., M.G.M., J.M., T.N., Q.P.-H., C.S.P., R.S., S.G.S., H.S., P.T., S.M.W., S.Z., P.A., C.J.C., M.M., B.F.F.O., K.W., H.Y., A.B., A.N.B., J.G., G.R., R.F.S., O.S. and Z.Z. (equal contributions by C.C., N.R.D., D.D., N.A.F., Y.H., A.K., K.-V.L., F.L., Y.S., C.M.S. and L.U.; jointly supervised and contributed by A.B., A.N.B., J.G., G.R., R.F.S., O.S. and Z.Z.). Data collection and coordination were carried out by N.A.F., A.K., K.-V.L., J.Z. M.D.P., Q.X., C.Y., K.A.H., P.B., R.S., S.G.S., B.F.F., A.B., G.R. and A.N.B. (equal contributions by N.A.F., A.K., K.-V.L., J.Z. M.D.P. and Q.X.; jointly supervised by A.B., G.R. and A.N.B.). Processing of RNA-seq data was carried out by N.A.F., A.K., K.-V.L., C.J.C., S.G.S., A.N.B., A.B. and G.R. (equal contributions by N.A.F., A.K. and K.-V.L.; jointly supervised by A.N.B., A.B. and G.R.). Analyses of eQTLs were carried out by C.C., K.-V.L., N.A.F., A.K., L.U., H.K., S.M.W., J.O.K., A.B., R.F.S., G.R. and O.S. (equal contributions by C.C. and K.-V.L.; jointly supervised by A.B., R.F.S., G.R. and O.S.). Analyses of allelic expression were carried out by L.U., F.L., H.K., J.M., S.E., M.R.H., Z.Z., O.S. and R.F.S. (equal contributions by L.U. and F.L.; jointly supervised by Z.Z., O.S. and R.F.S.). Analyses of alternative splicing were carried out by A.K., Y.S., C.M.S., K.-V.L., S.G.S., M.G.M., G.R. and A.N.B. (equal contributions by A.K., Y.S. and C.M.S.; jointly supervised by G.R. and A.N.B.). Analyses of alternative promoters were carried out by D.D., T.N., C.C., K.-V.L., P.T. and J.G. Analyses of fusions were carried out by N.A.F., Y.H., L.G., A.B. and Z.Z. (equal contributions by N.A.F. and Y.H.; jointly supervised by A.B. and Z.Z.). Analyses of RNA editing were carried out by D.L., S.L., H.S., Y.H., S.Z., Q.P.-H., H.Y. and K.W. (equal contributions by D.L. and S.L.; jointly supervised by H.Y. and K.W.). Mutational signature analysis was carried out by L.U., S.M.W., K.-V.L., R.F.S. and O.S. (jointly supervised by R.F.S. and O.S.). Meta-analyses of transcriptome alterations were carried out by N.R.D., F.L., K.-V.L., F.Z., D.D., N.A.F., A.K., S.L., R.F.S., H.S., R.S., Y.H., S.G.S., A.B., A.N.B., Z.Z. and G.R. (jointly supervised by A.B., A.N.B., Z.Z. and G.R.). A.B., G.R. and A.N.B. coordinated the overall project as working group leaders. Writing was carried out by C.C., N.R.D., D.D., N.A.F., Y.H., A.K., K.-V.L., F.L., Y.S., C.M.S., L.U., A.B., A.N.B., J.G., G.R., R.F.S., O.S. and Z.Z. (equal contributions by C.C., N.R.D., D.D., N.A.F., Y.H., A.K., K.-V.L., F.L., Y.S., C.M.S. and L.U.; jointly supervised and contributed by A.B., A.N.B., J.G., G.R., R.F.S., O.S. and Z.Z.) with input from all other co-authors.

**Competing interests** M.M. is a scientific advisory board chair of, and consultant for, Origimed, receives research funding from Bayer and Ono Pharma, and has patent royalties from LabCorp. G.R. is on the scientific advisory board of Computomics GmbH and receives research funding from Roche Diagnostics and Google. R.S. received honorariums for speaking at meeting organized by Roche and AstraZeneca. All the other authors have no competing interests.

#### Additional information

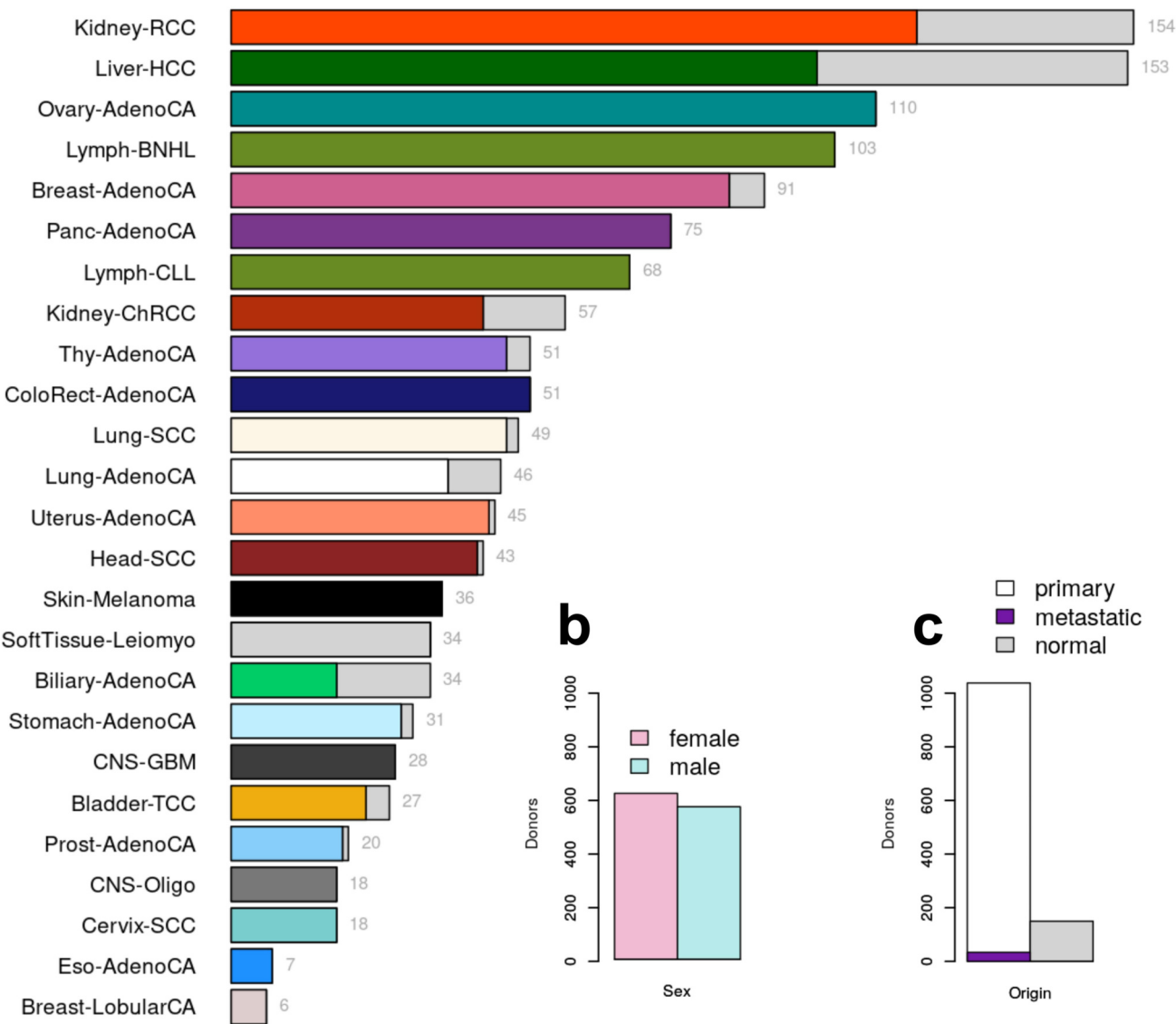
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1970-0>.

**Correspondence and requests for materials** should be addressed to A.B., A.N.B. or G.R.

**Peer review information** *Nature* thanks Nicolas Robine and the other anonymous reviewer(s) for their contribution to the peer review of this work.

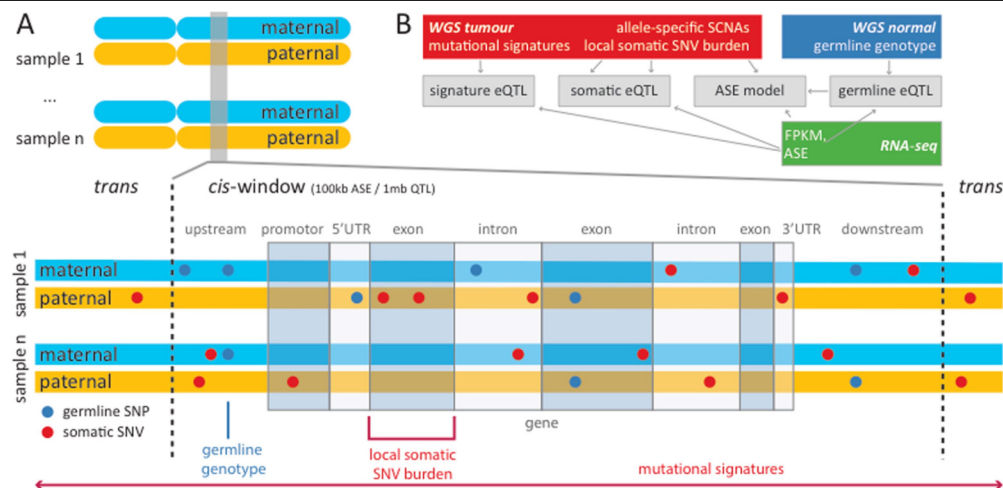
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

a



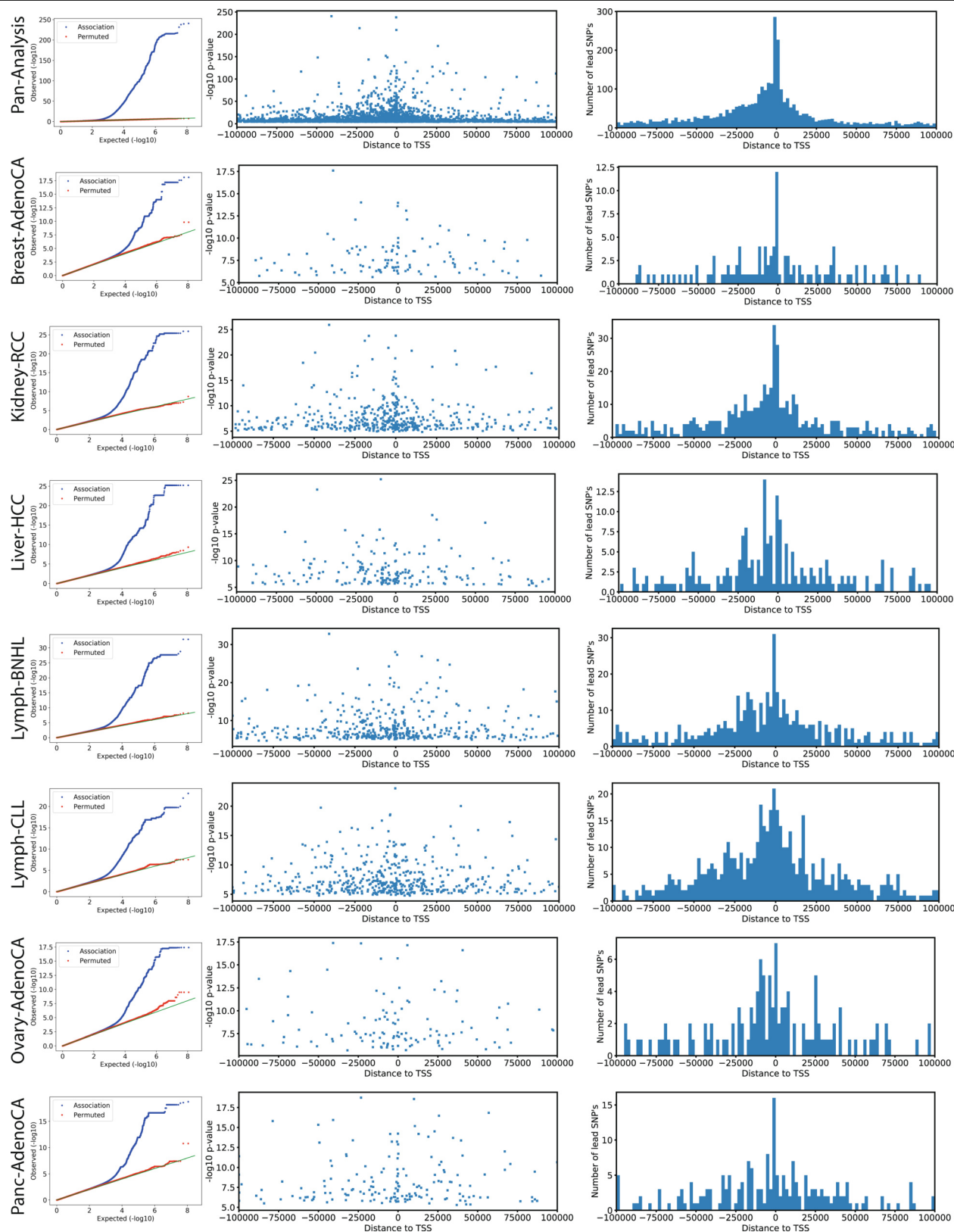
**Extended Data Fig. 1|Pan-cancer expression profiling of 1,188 PCAWG donors. a,** Tumour and normal RNA-seq data from 27 histotypes. The total number of samples is shown to the right of the bars. Grey bars denote matched

healthy samples. **b,** Number of female versus male donors. **c,** Total number of tumour and matched healthy samples from the PCAWG study. A subset of tumours (dark violet) was metastatic.



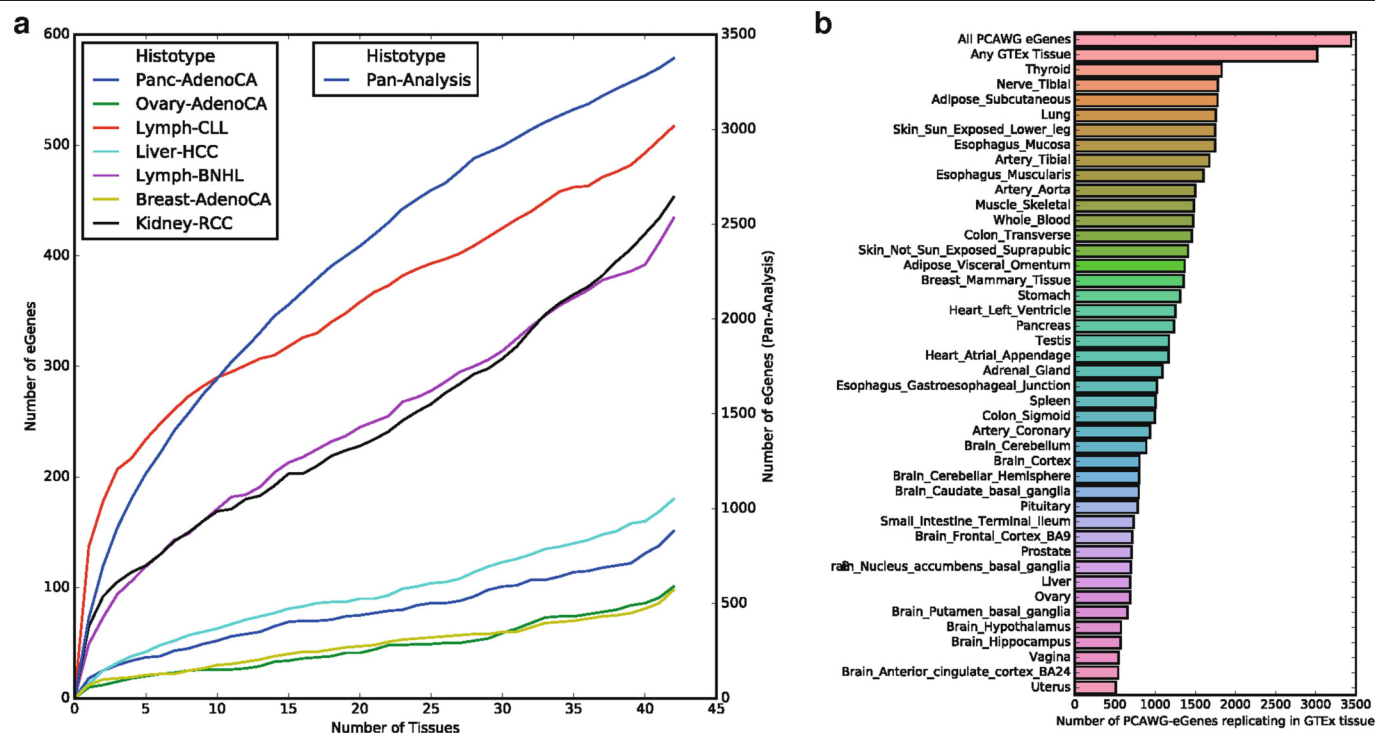
**Extended Data Fig. 2 | Overview of the different sources of genetic variation considered in the analysis. a,** For analyses of *cis* regulation, mono-allelic single-nucleotide germline variants (single nucleotide polymorphisms (SNPs), blue) were individually tested for association with total gene expression using standard eQTL approaches. Owing to their low recurrence in the cohort, somatic SNVs were aggregated in burden categories depending on their position relative to the gene tested (for example, promoter, 5' UTR or intron). Local SNV burdens were then tested for association with ASE globally across all genes, as well as with total expression on a per-gene level using eQTL approaches. *Trans* effects were estimated by testing total gene expression for

association with mutational and epigenetic signatures. Window sizes were 1 Mb for all somatic *cis*-eQTL analyses, and 100 kb for ASE and germline *cis*-eQTL. **b,** Overview of the different datasets and their contributions to the analyses described in **a**. Germline genotypes were derived from the matched healthy whole-genome sequencing (WGS) samples. Allele-specific SCNAs, mutational signatures and local SNV burdens were derived from the tumour WGS in comparison to the unaffected WGS samples. ASE and total expression (FPKM) were derived from the tumour and normal RNA-seq data. Arrows indicate dependencies between individual analyses carried out.



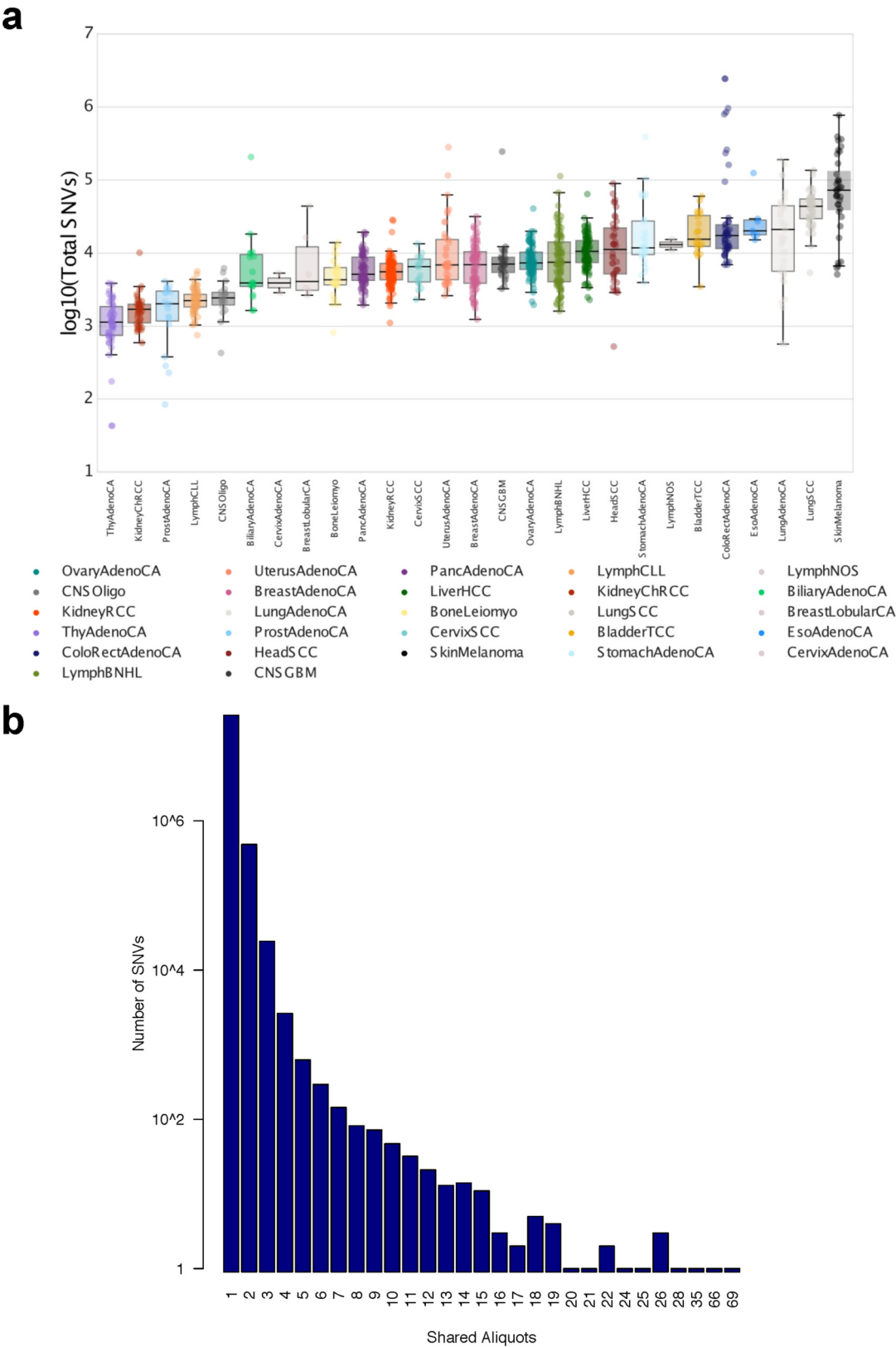
**Extended Data Fig. 3 | Germline eQTL lead variants.** Left, quantile–quantile (Q–Q) plot of  $P$  values of germline eQTL lead variants in the pan-cancer and histotype-specific analysis (FDR  $\leq 5\%$ , blue) and  $P$  values of the same analysis

after permutation (random permutation of patients, red). Middle and right, distributions of distance to the respective TSS of all germline eQTL lead variants in the pan-cancer and histotype-specific analysis.

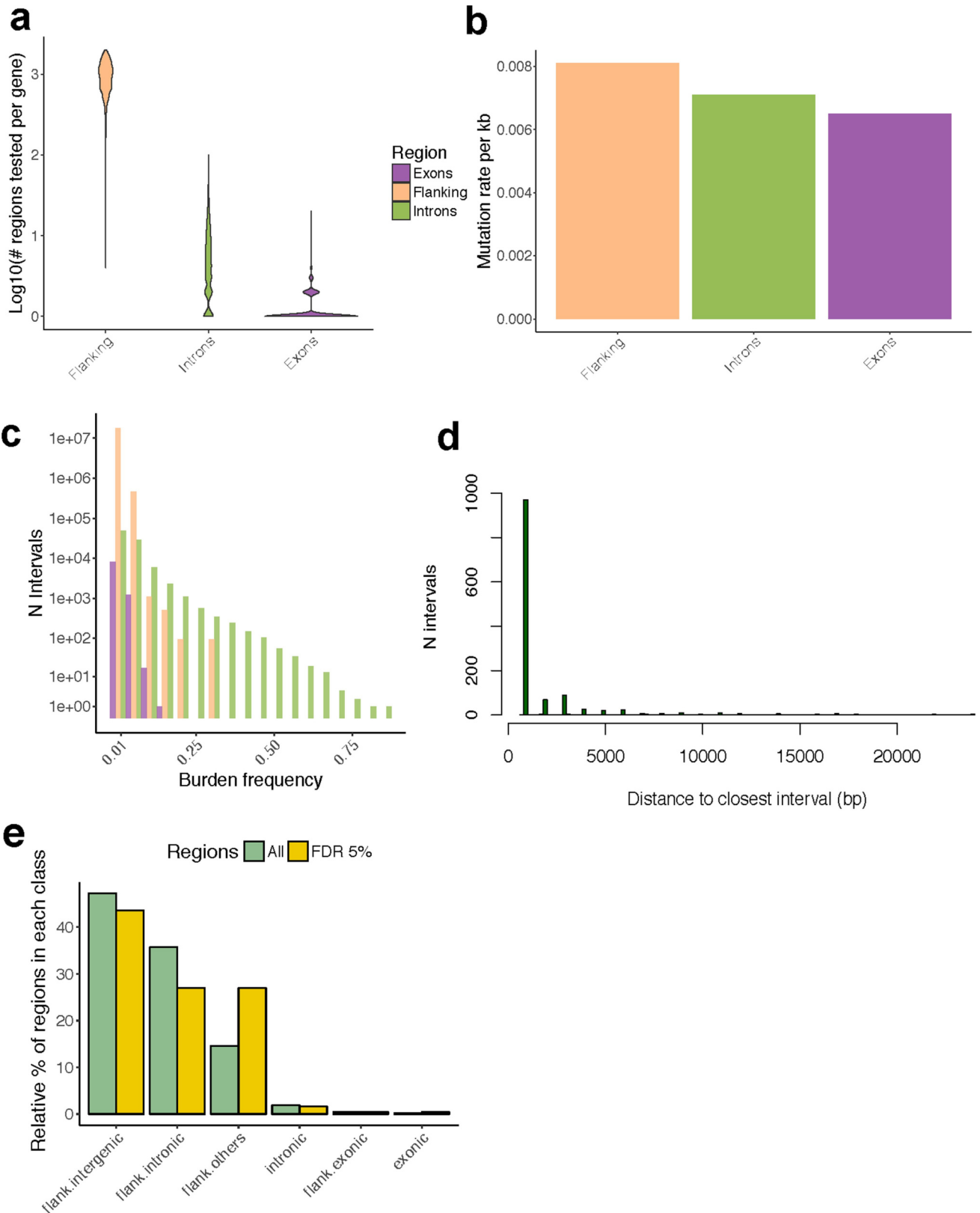


**Extended Data Fig. 4 | PCAWG-specific eGenes. a**, Number of PCAWG-specific eGenes in relation to eQTL replication in various numbers of GTEx tissues. **b**, Number of eGenes of the PCAWG pan-analysis replicating in corresponding GTEx tissues.



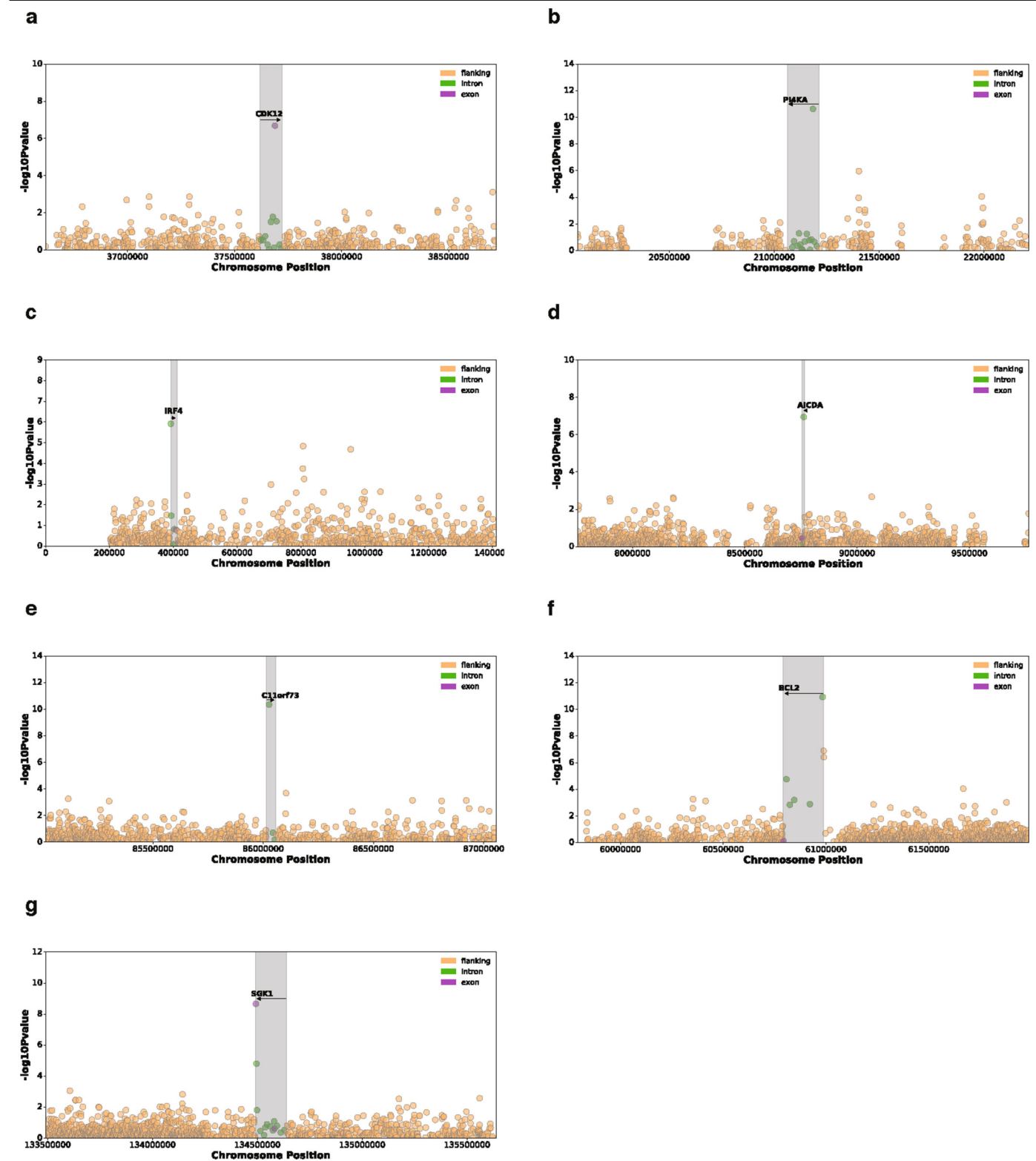


**Extended Data Fig. 5 | *Cis*-mutational somatic burden. a**, Total number of somatic mutational load per cancer type. Median numbers of SNVs range from 1,139 in thyroid adenocarcinoma to 72,804 in skin melanoma. **b**, Number of recurrent somatic SNVs shared by increasing numbers of patients. A small fraction of 86 SNVs is detected in more than 1% of the cohort (12 patients).



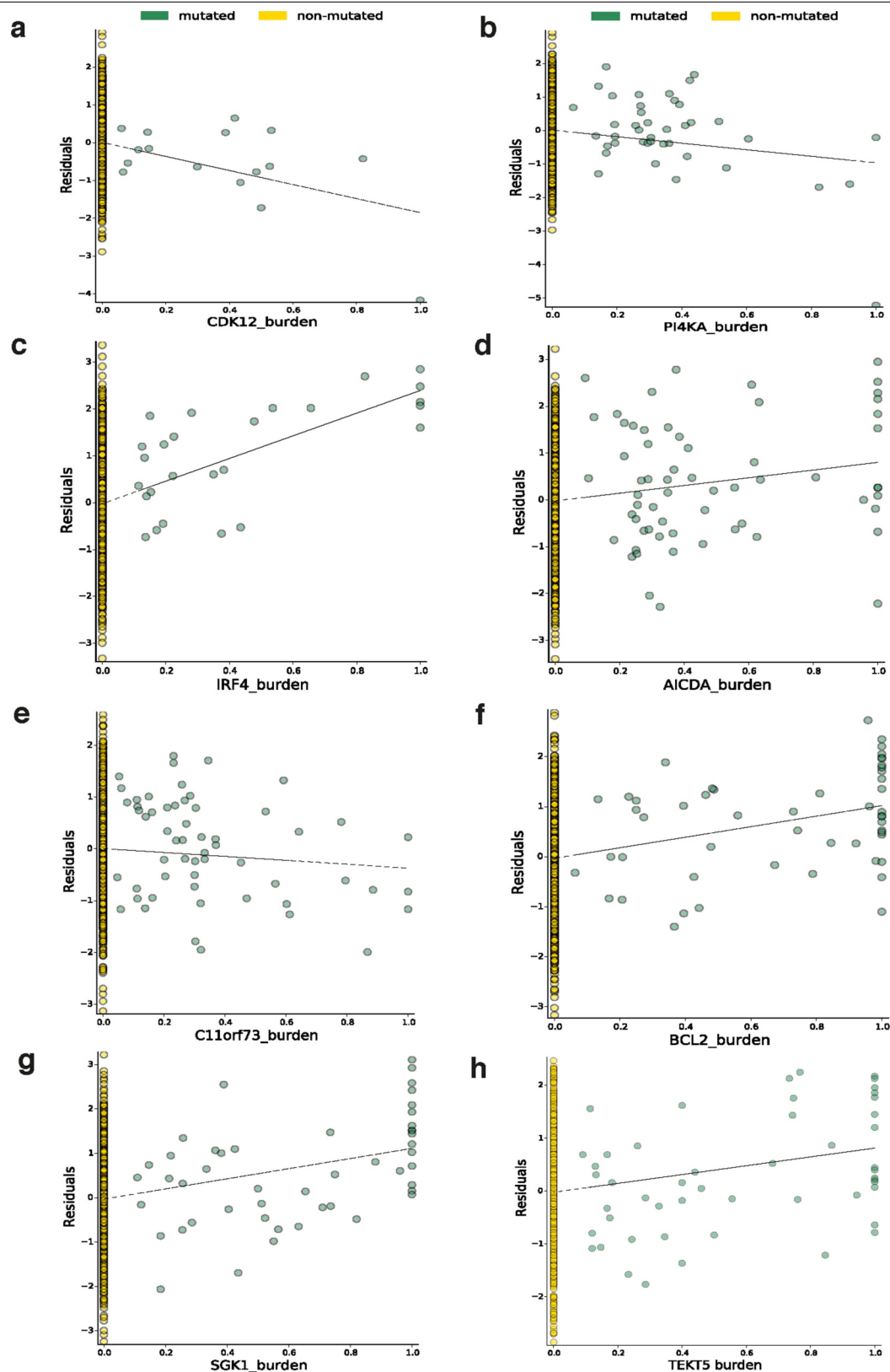
**Extended Data Fig. 6 | Somatic mutation rate and burden frequency by type of region tested.** **a**, Number of mutated regions tested per gene with somatic burden frequency  $\geq 1\%$ . **b**, Mutation rate per kilobase. **c**, Burden frequency, stratified by the type of interval tested (flanking, exonic or intronic). **d**, Distribution of distances (bp) of the leading intervals ( $FDR \leq 5\%$ ) to the closest (left and right) interval such that the association  $P$  value decreases by at least one order of magnitude (99% of the distribution is shown). **e**, Breakdown of all genomic regions tested ( $n = 1,049,102$  with burden frequency  $\geq 1\%$ ) and of

the 567 genomic regions that underlie the observed somatic *cis*-eQTL at a FDR of 5% (intronic denotes eGene intron; exonic denotes eGene exon; flank. denotes 2-kb flanking region within 1 Mb distance to the eGene start and end; flank.intergenic denotes flanking region in a genomic location without gene annotations; flank.intronic denotes flanking region overlapping an intron of a nearby gene; and flank.others denotes flanking region partially overlapping several annotations of a nearby gene).

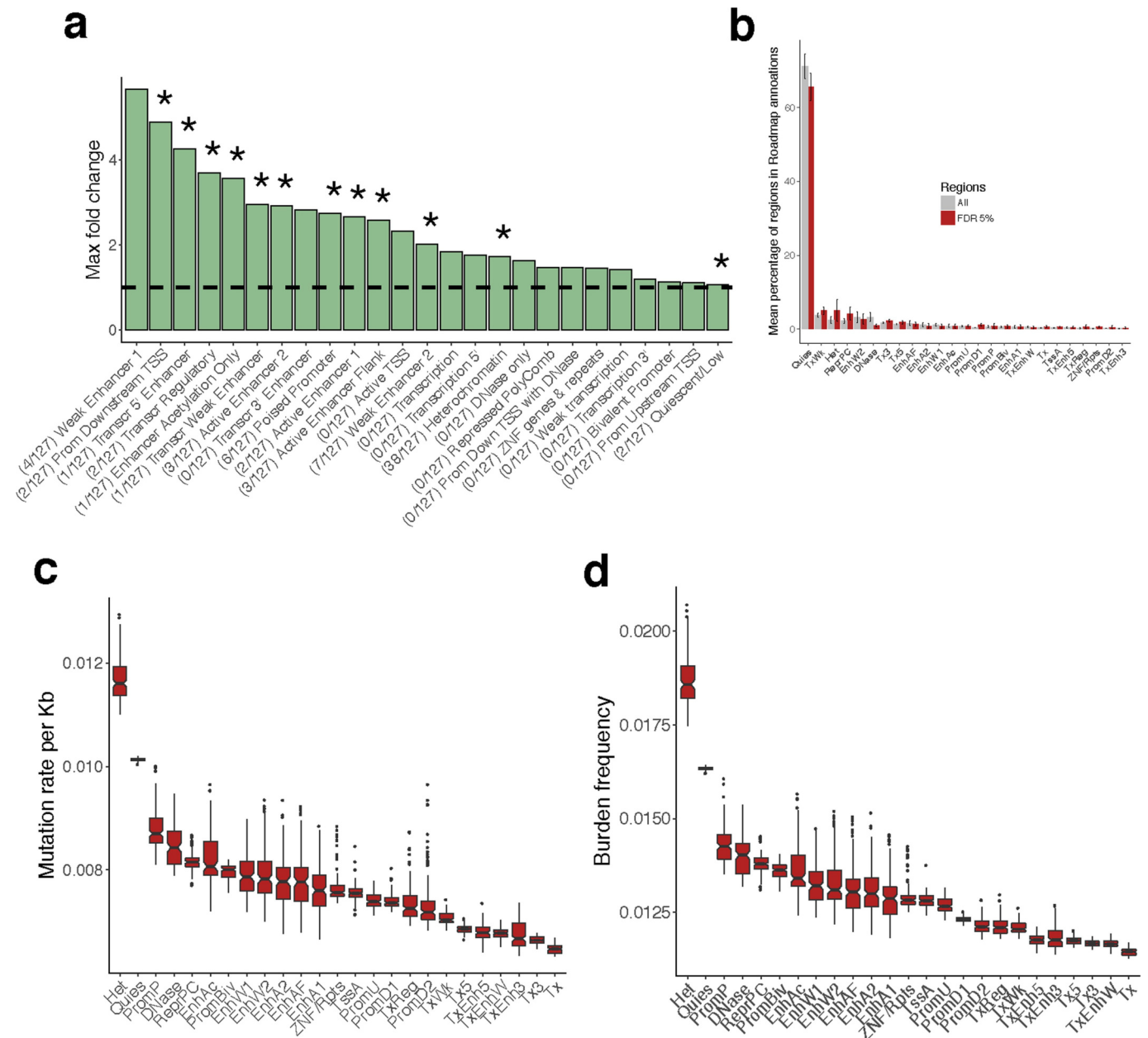


**Extended Data Fig. 7 | Manhattan plots of seven somatic eGenes associated with genic lead burden.** Altogether, 11 genic somatic eQTLs showed significant changes in gene expression associated with somatic burdens within the gene

boundaries (intronic or exonic). The seven genes shown here are known to be important in the pathogenesis of specific cancers. **a**, *CDK12*. **b**, *PI4KA*. **c**, *IRF4*. **d**, *AICDA*. **e**, *C11orf73* (also known as *HIKESHI*). **f**, *BCL2*. **g**, *SGK1*.



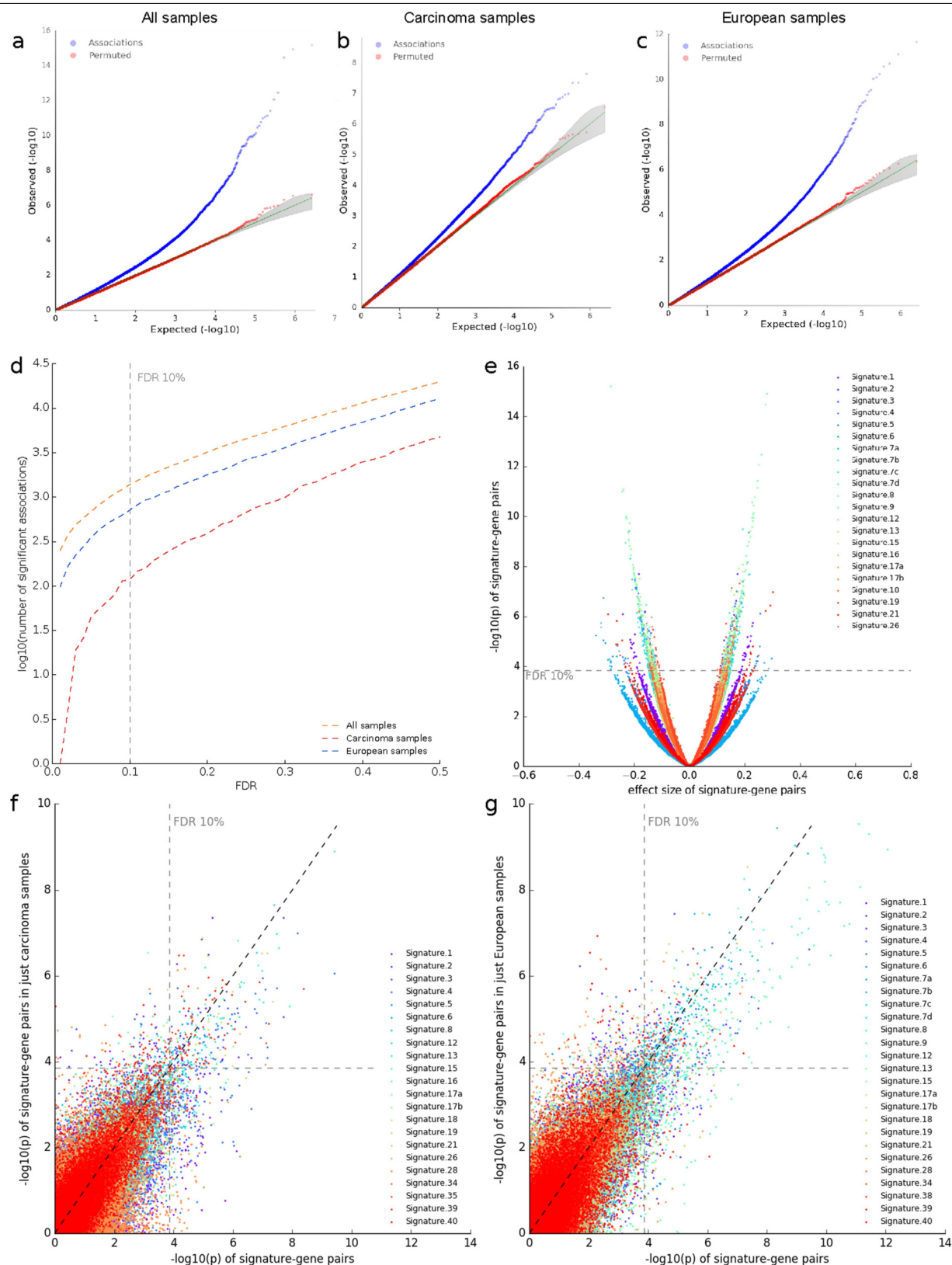
**Extended Data Fig. 8 | Scatter plots of eight somatic eGenes.** Plots show the effect of the lead weighted burden on the gene expression residuals (obtained as described in the Methods) of these genes. **a**, *CDK12*. **b**, *PI4KA*. **c**, *IRF4*. **d**, *AICDA*. **e**, *C11orf73*. **f**, *BCL2*. **g**, *SGK1*. **h**, *TEK5*.



**Extended Data Fig. 9 | Roadmap epigenome marks overlapping flanking intervals with somatic burden.** **a**, Maximum fold enrichment of epigenetic marks from the Roadmap Epigenomics Project across 127 cell lines. The number of cell lines with significant enrichments is indicated in parentheses ( $FDR \leq 10\%$ ); asterisks denote significant enrichments in at least one cell line. **b**, Mean percentages (over the 127 cell lines) of regions overlapping (by at least 10% of their length) Roadmap epigenome marks, calculated using all genomic flanking regions ( $n = 1,637,638$ ) and the subset of 556 flanking intervals associated with somatic eQTL ( $FDR \leq 5\%$ ). **c**, Mutation rate per kilobase. **d**, Burden frequency (across the 127 cell lines) of the 556 flanking intervals in

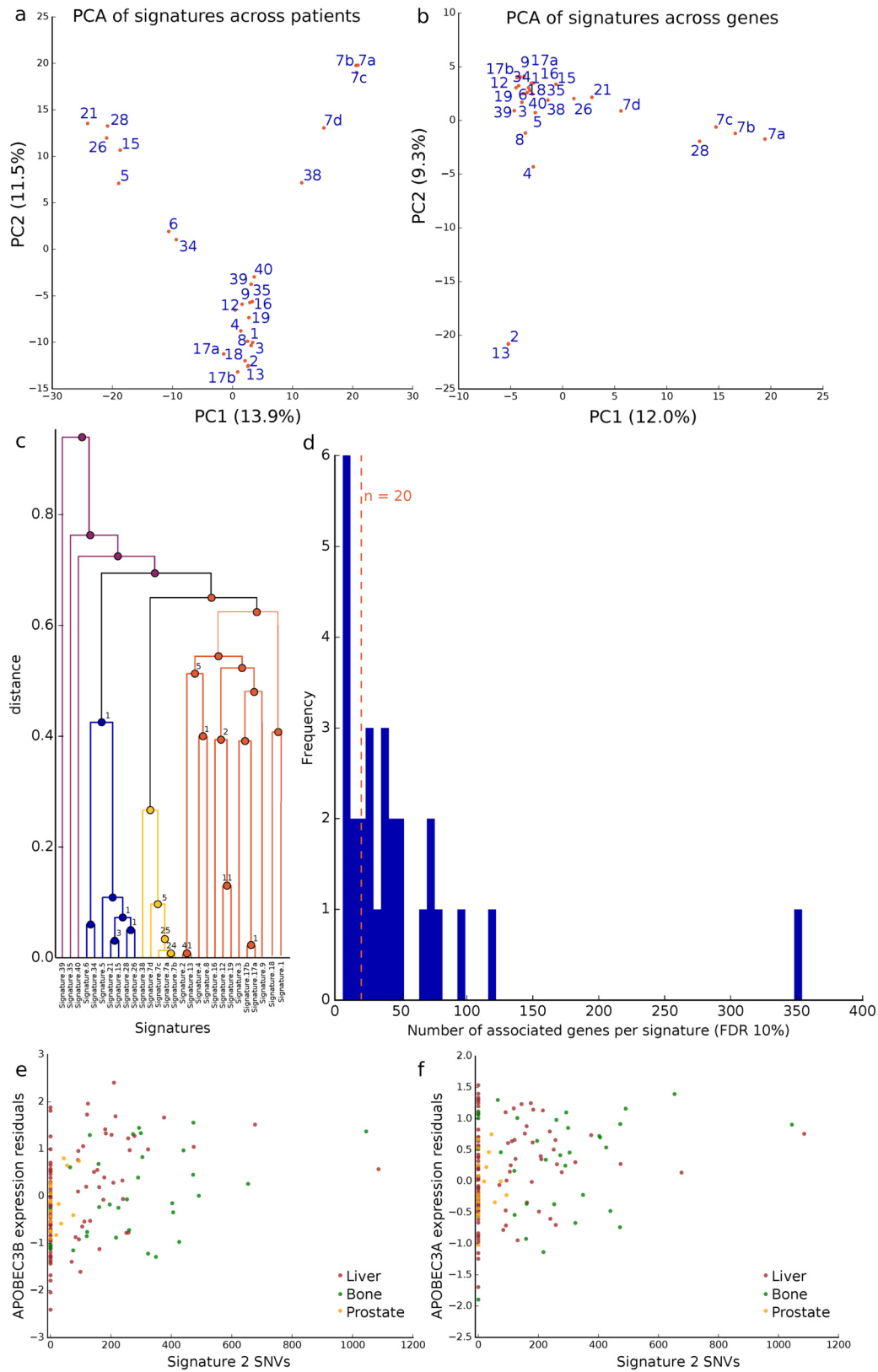
somatic eQTLs ( $FDR \leq 5\%$ ), overlapping 25 Roadmap epigenome marks. DNase, DNase only; EnhA, active enhancer; EnhAc, enhancer acetylation only; EnhAF, active enhancer flank; EnhW, weak enhancer; Het, heterochromatin; PromBiv, bivalent promoters; PromD, promoter downstream; PromP, poised promoters; PromU, promoter upstream; Quies, quiescent/low; ReprPC, repressed PolyComb; TssA, active TSS; TxReg, transcription regulatory; ZNF/Rpts, ZNF genes and repeats; Tx, transcription; Tx3, transcription 3'; Tx5, transcription 5'; TxEnh3, transcription 3' enhancer; TxEnh5, transcription 5' enhancer; TxEnhW, transcription weak enhancer; TxWk, weak transcription.





**Extended Data Fig. 10 | Quality control of the association studies between gene expression and mutational signatures.** **a–c**, Q–Q plots of the  $P$  values of the linear model to associate expression of 18,831 genes with 28 mutational signatures across all 1,159 patients (**a**), 877 patients with carcinoma (**b**), or 891 European patients (**c**). **d**, Number of significant associations ( $\log_{10}$ -transformed) at different FDR thresholds (across all patients, patients

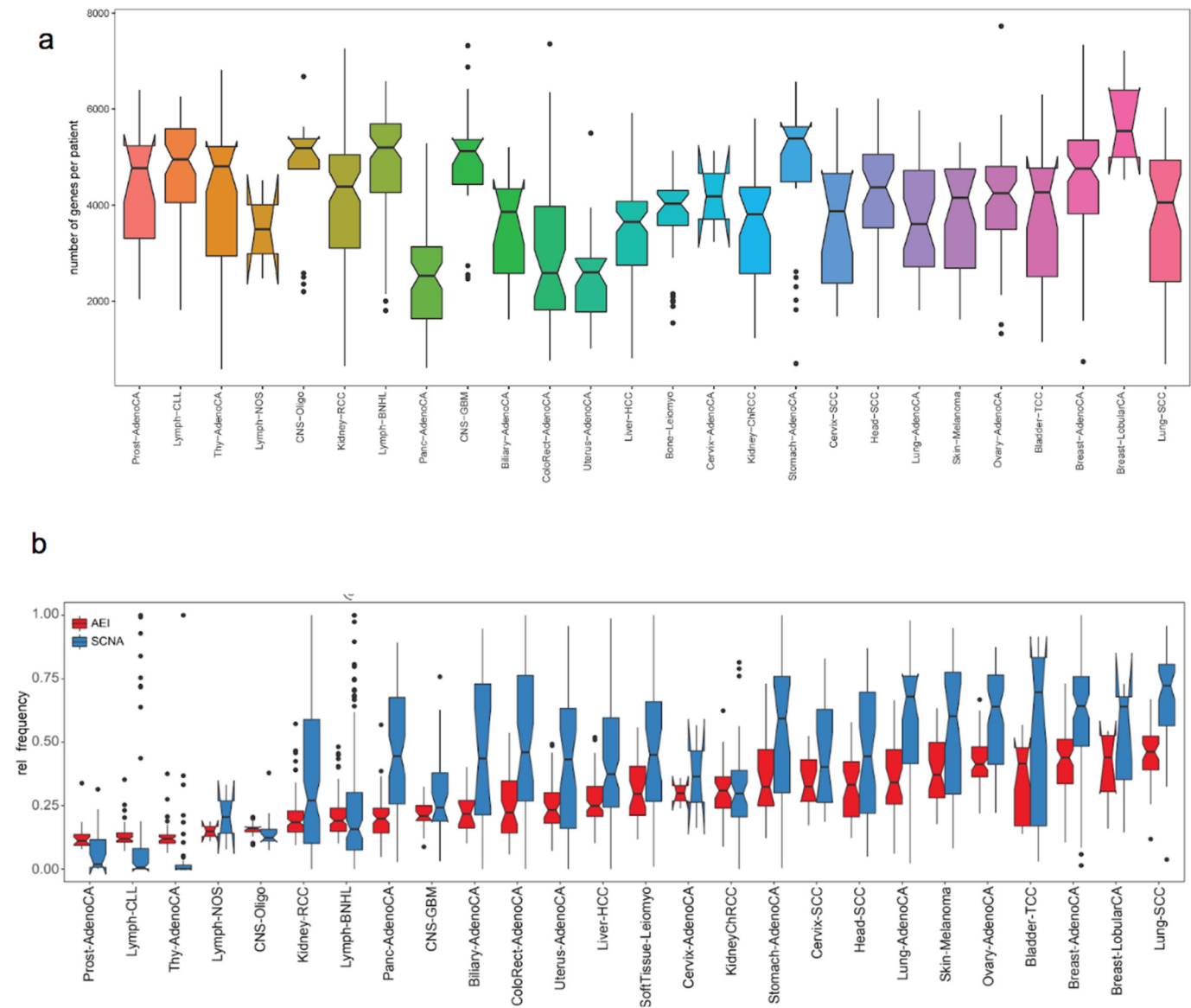
with carcinoma and European patients). **e**, Volcano plot of directionality of effects in the analysis of all patients. **f**, **g**, Comparison of analyses between all patients and patients with carcinoma (**f**) and between all patients and European patients (**g**). The  $-\log_{10}(P)$  values per signature–gene pair are correlated ( $r = 0.763$  (**f**) and  $r = 0.789$  (**g**), Pearson correlation coefficient), especially above an FDR threshold of 10%.



**Extended Data Fig. 11** | See next page for caption.

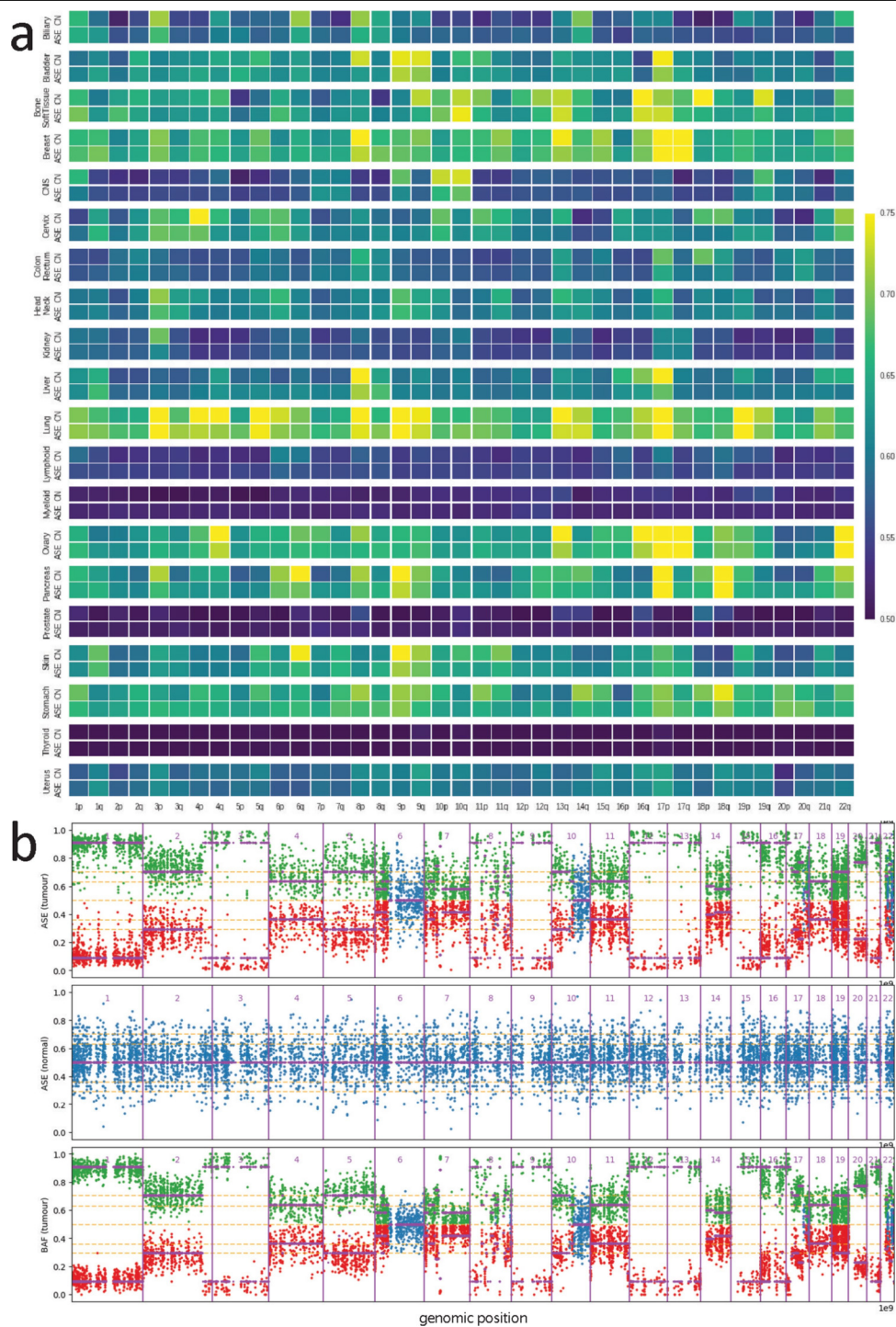
**Extended Data Fig. 11 | Relationship between mutational signatures and gene expression patterns. a, b,** Principal component analysis (PCA) of signatures across 1,159 patients (PCA on signature-specific SNVs per patient) (**a**) and signature–gene expression associations across 18,831 genes (PCA on adjusted *P* values of signature–gene expression associations) (**b**). The PCA on the SNVs recapitulates known interdependencies, for example, between signatures 7, whereas the PCA on the signature–gene association studies also emphasizes functional relatedness, for example, between signatures 2 and 13. **c,** Hierarchical clustering of signatures. The numbers at the nodes indicate the number of genes commonly associated with two to four respective signatures. The dendrogram shows genes that are associated with more than one signature mostly owing to similar SNV patterns of these signatures across patients.

**d,** Frequency of number of significantly associated genes per signature ( $\text{FDR} \leq 10\%$ ). Although many signatures are significantly associated with a few genes, 18 signatures are associated with more than 20 genes. Signature 9 is associated with more than 350 genes. Vice versa, 1,009 genes are associated with only one signature, 129 with two, 32 with three, 5 with four and 1 with five signatures. **e, f,** Mutational signature–gene associations, depicting positive associations between the expression of the canonical APOBEC pathway genes *APOBEC3B* (**e**) and *APOBEC3A* (**f**) and signature 2. The associations within the three cancer type with the strongest correlation between signature and gene expression (hepatocellular carcinoma (Liver–HCC), bone leiomyosarcoma (Bone–Leiomyo) and prostate adenocarcinoma (Prost–AdenoCA)) are shown.



**Extended Data Fig. 12 | ASE analysis. a.** All types of cancer are ordered by the average AEI frequency. The numbers of genes per patient for which ASE could be quantified are shown, stratified according to cancer type, resulting in between 588 and 7,728 genes per patient. **b.** Distribution of the fraction of

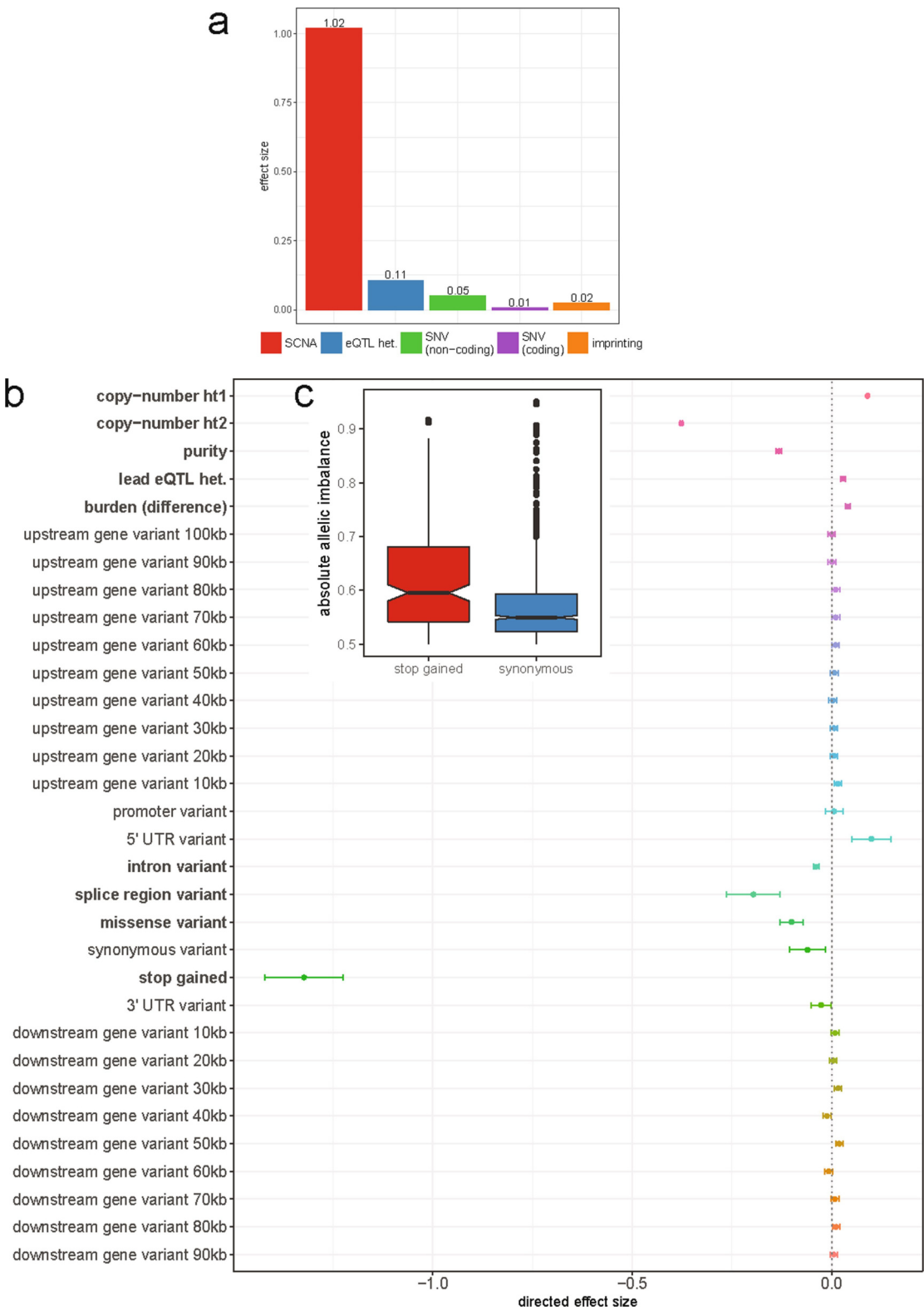
genes with AEI (red) and SCNAs (blue) over the number of measurable genes for each patient across the cohort. Cancer types with high chromosomal instability also exhibit highest amounts of AEI.



**Extended Data Fig. 13 | SCNAs as major driver for allelic dysregulation in cancer. a.** Absolute allelic expression closely follows allelic imbalance at the genomic level. Values of 0.5 (blue) denote equal number of reads from both alleles. Values of 1 (yellow) reflect mono-allelic expression or

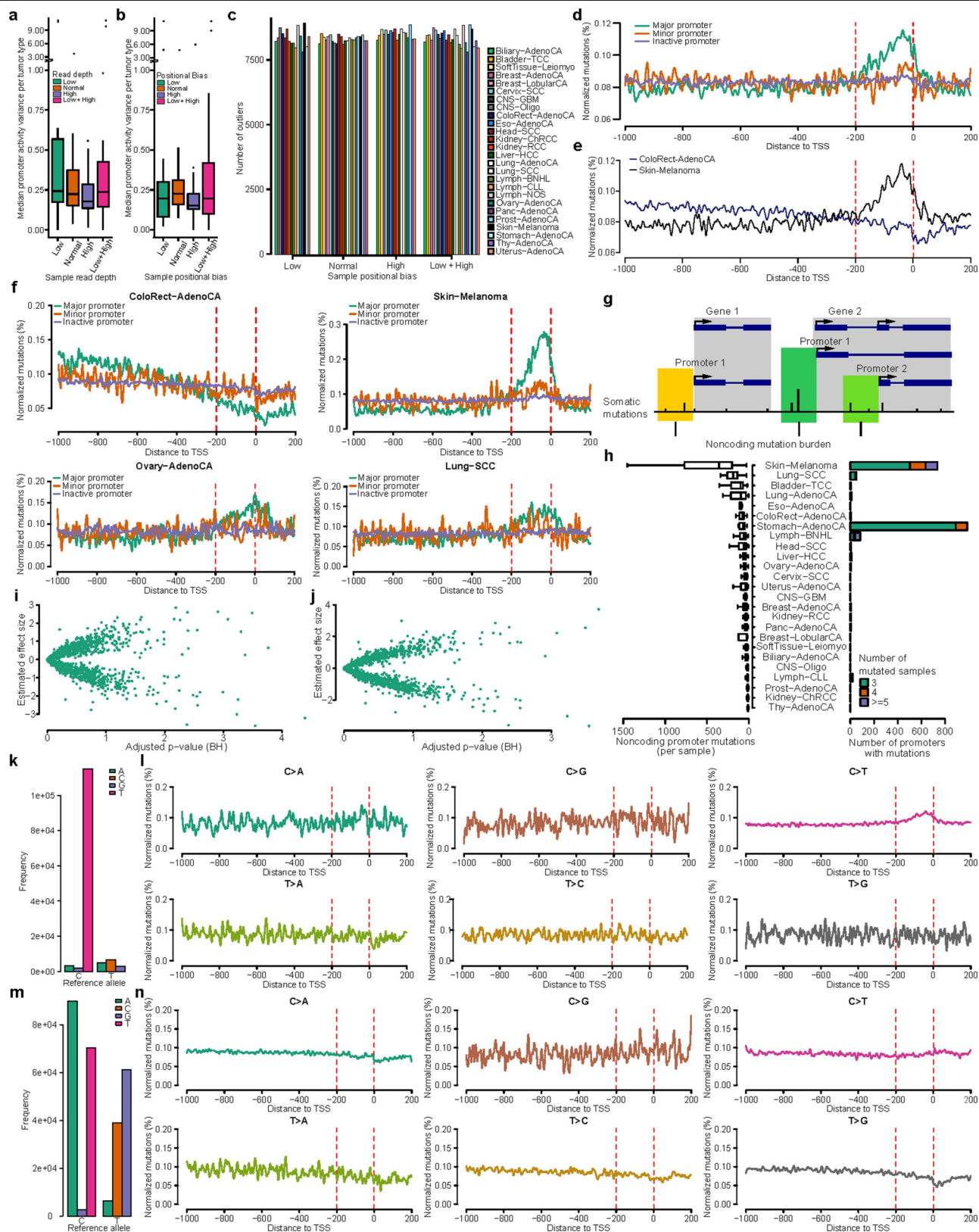
regions with loss of heterozygosity. **b.** Comparison between B-allele frequency (BAF) and ASE ratios from a single patient with lung cancer (LUAD-US) with profound chromosomal instability shows strong correlation between allelic imbalance on expression and genomic levels.





**Extended Data Fig. 14 | Determinants of AEI.** **a**, Standardized effect sizes on the presence of AEI, taking only SCNAs, germline eQTLs, coding and non-coding mutations into account. In summary, SCNAs accounted for 86.1% of the total effect size, followed by germline eQTLs (9.0%) and somatic SNVs (4.8%). **b**, Relevance of individual somatic mutation types ('copy-number ht1' and

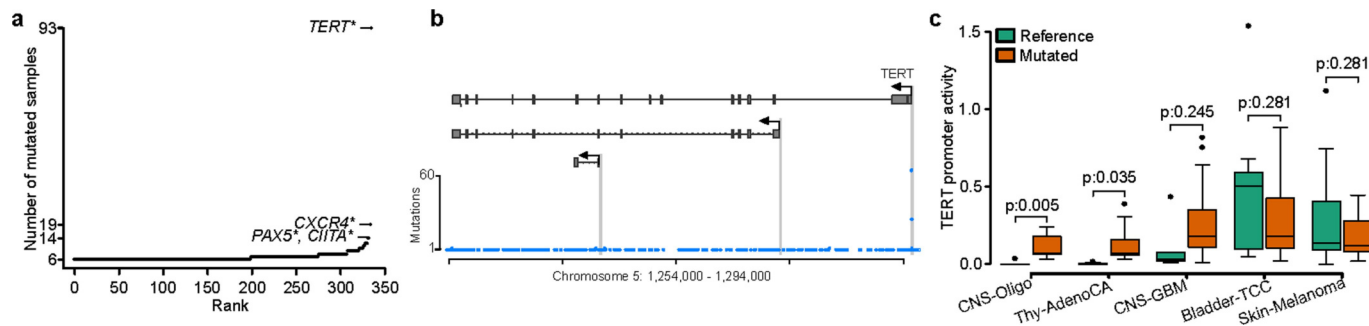
'copy-number ht2' as local allele-specific SCNAs of haplotypes 1 and 2, respectively), germline eQTLs and other covariates for the ASE ratio. Significant covariates ( $FDR \leq 5\%$ ) are highlighted in bold. **c**, Comparison of the effect of protein-truncating variants (stop-gained) and synonymous variants on the ASE ratio.



**Extended Data Fig. 15** | See next page for caption.

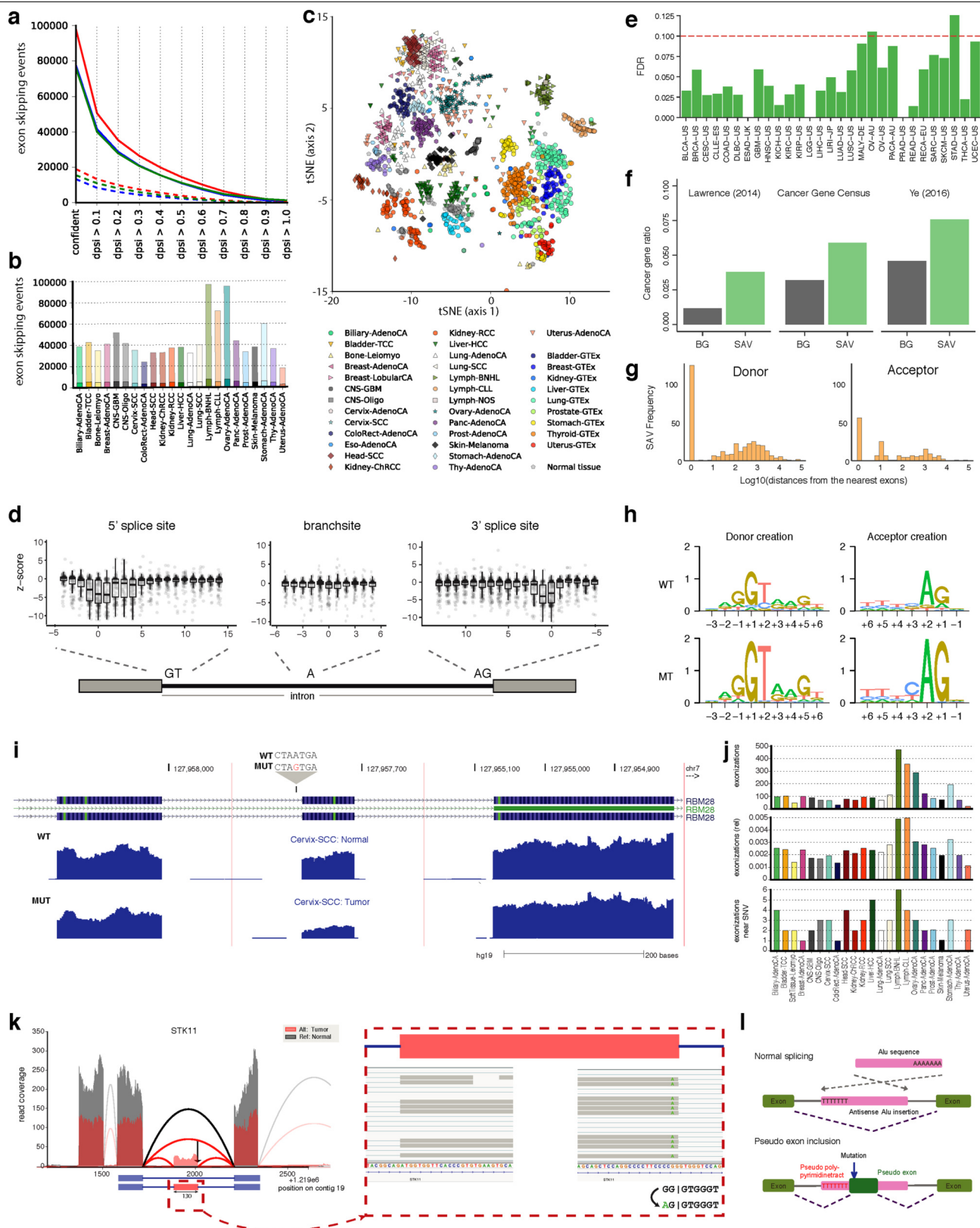
**Extended Data Fig. 15 | Overview of estimations of promoter activity and non-coding promoter mutations associations and patterns.** **a, b**, The technical variation of the promoter activity estimates across varying library depth (**a**) and positional bias (**b**). **c**, The number of outlier promoters per tumour type according to promoter activity variance (variance larger than  $1.5 \times$  the interquartile range). **d**, Distribution of promoter mutations around promoters across the PCAWG cohort for major, minor and inactive promoters. Red lines indicate the window 200-bp upstream of a TSS, in which major promoters show an enrichment of mutations whereas minor and inactive promoters do not. **e**, Distribution of promoter mutations around promoters for the top two most mutated types of cancer (skin melanoma and colorectal adenocarcinoma (ColoRect-AdenoCA)). Colorectal adenocarcinoma displays a very different mutational pattern from other types of cancer. **f**, Distribution of promoter mutations around major, minor and inactive promoters across several types of cancer. Red lines indicate the window 200-bp upstream of a

TSS, in which major promoters show an enrichment of mutations whereas minor and inactive promoters do not. **g**, Schematic of the calculation of non-coding promoter mutational burden. **h**, Overview of non-coding promoter mutations per sample and the number of mutated promoters per tumour type for promoters with at least three mutated samples. **i, j**, Association of absolute (**i**) and relative (**j**) promoter activity with promoter mutations across all samples. **k, l**, Overview of promoter mutations for skin melanoma tumours. **k**, Most promoter mutations are C>T, which indicates UV-induced DNA damage. **l**, Distribution of promoter mutations for each mutation class reveals the enrichment of C>T mutations around the 200-bp window upstream. **m, n**, Overview of promoter mutations for colorectal adenocarcinoma tumours. **m**, Most promoter mutations are C>A and C>T. **n**, Distribution of promoter mutations for each mutation class does not display an enrichment of mutations around the 200-bp window upstream, differing from the mutation pattern of skin melanoma tumours.



**Extended Data Fig. 16 | *TERT* promoter mutations.** **a**, Promoters ranked by the number of mutated samples across all types of cancer in a 200-bp window. Asterisk indicates cancer census genes. **b**, The *TERT* locus and number of mutations observed at each position. The first promoter shows a highly

recurrent non-coding mutation reported previously<sup>118,119</sup>. **c**, Comparison of *TERT* promoter activity for mutated and non-mutated samples per tumour type.

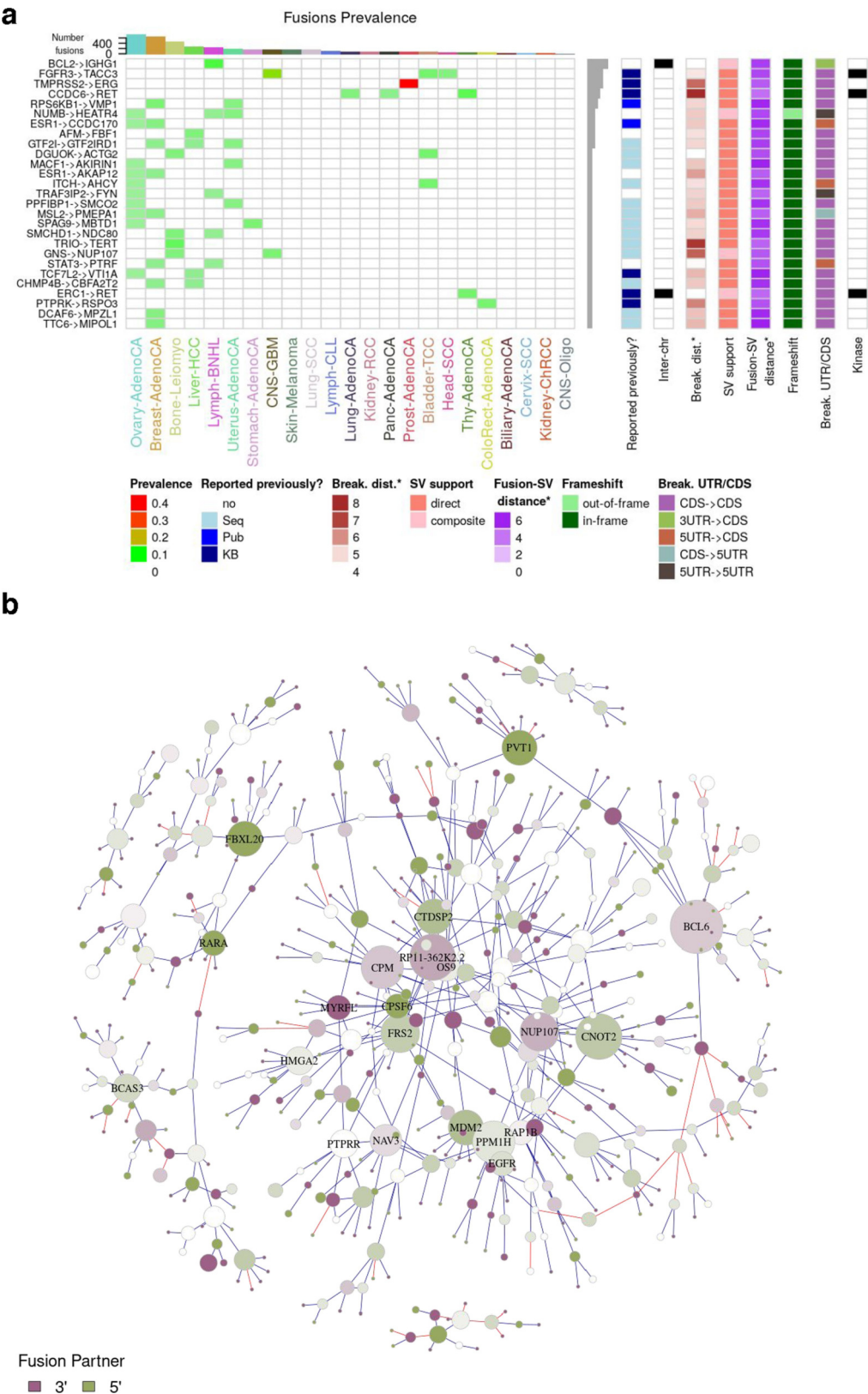


Extended Data Fig. 17 | See next page for caption.



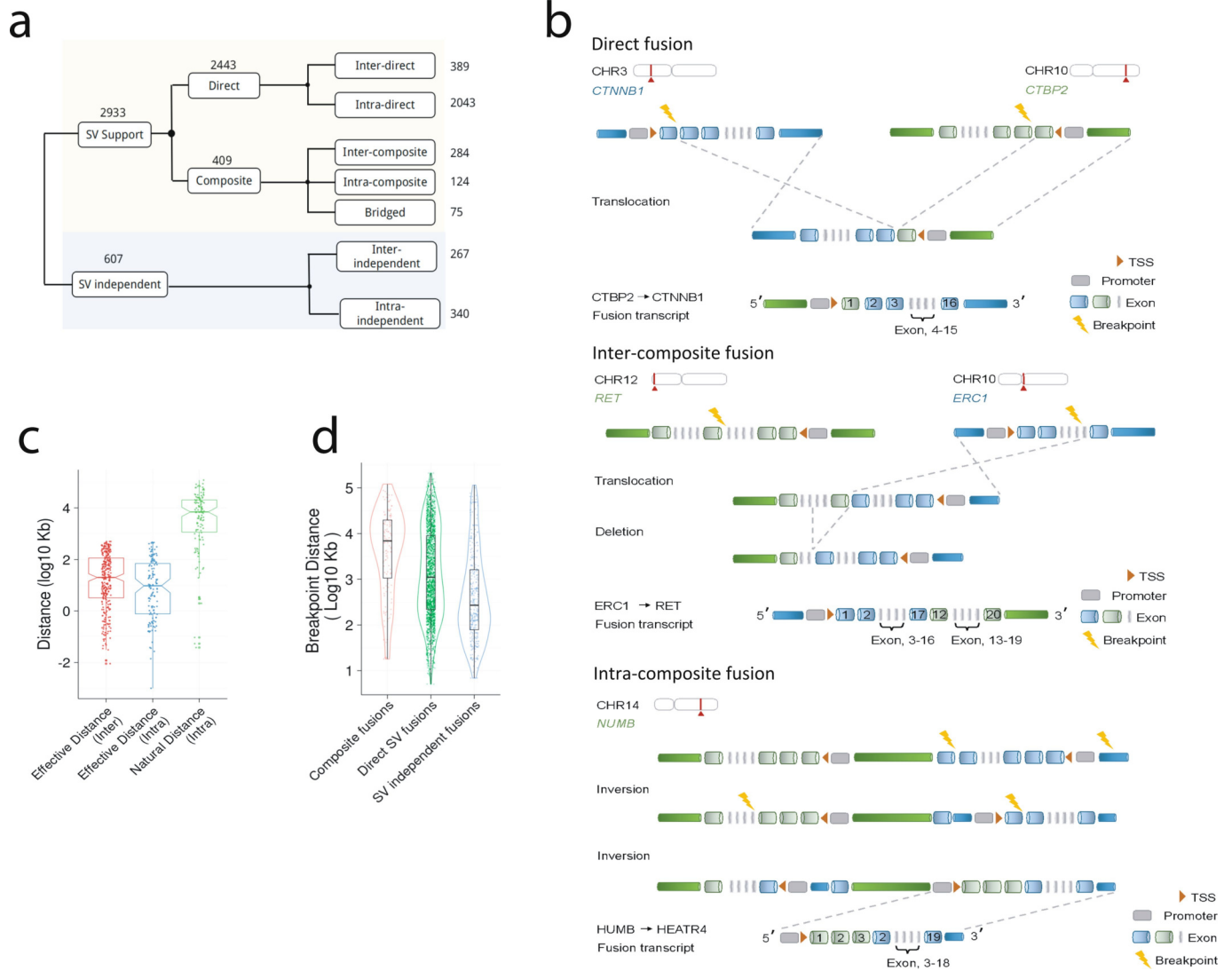
**Extended Data Fig. 17 | Alternative splicing and association with somatic mutations.** **a**, Number of exon-skipping events confirmed at different  $\Delta$ PSI thresholds in tumour (red), matched healthy (green) and GTEx (blue) samples for liver tissue. Dashed lines show the subset of exon-skipping events that only contain annotated introns. **b**, Number of exon-skipping events confirmed at a  $\Delta$ PSI level of greater than 0.3 for the individual histotypes. Transparent section of bars represents the fraction of novel events, containing at least one unannotated intron. **c**, Splicing landscape for exon-skipping events. *t*-SNE analysis based on exon-skipping PSI values for all ICGC tumour and healthy samples together with tissue-matched GTEx samples. **d**, Position-specific effect of somatic mutations on alternative splicing. Magnitude and direction of mutation-associated splicing alterations. **e**, Permutation-based FDR values for SAV detection based on the different types of cancer. **f**, Cancer gene set enrichment for SAV sets, shown for cancer census gene set (middle) and sets determined in ref. <sup>48</sup> (left) and ref. <sup>120</sup> (right). **g**, Positional distributions (logarithms of distance from the nearest exons) of somatic variant creating novel splicing donors and acceptors. **h**, Sequence motif logos around somatic mutation creating novel splicing motifs. **i**, Example splicing effect of a branch-point mutation. UCSC genome browser RNA-seq coverage plots of cassette exon event in *RBM28* between mutant and wild type. Mutant (bottom track)

contains an A>G mutation 29 nucleotides upstream from the acceptor site of an affected exon. **j**, Distribution of new cassette exon events detected only within the PCAWG cohort. Top, number of events per histology type. Middle, events normalized to the total number of cassette exons detected in the histology types. Bottom, the number of exonization events per histotype for the subset with the novel cassette exons collocated to a somatic alteration near the acceptor or donor of the exon. **k**, Example of an exonization event in the tumour-suppressor gene *STK11*. RNA-seq read coverage for a part of the gene is shown in red for a donor carrying the alternate allele and in grey for a random donor with reference allele. The cassette exon event is shown as a schematic below, with blue (red) boxes denoting constitutive (alternative) exons and blue solid lines denoting introns. Magnified panels at the bottom show details from Integrative Genomics Viewer visualization, highlighting a somatic mutation at the 3' end of the cassette exon. The associated sequencing change is illustrated on the bottom right corner, in which the vertical bar denotes the exon-intron boundary. **l**, Alu-based exonization mechanism. Top, the presence of an Alu element in an intron in antisense alone will still result in normal splicing. Bottom, specific mutations of the Alu sequence creates new splice sites and results in exonization.



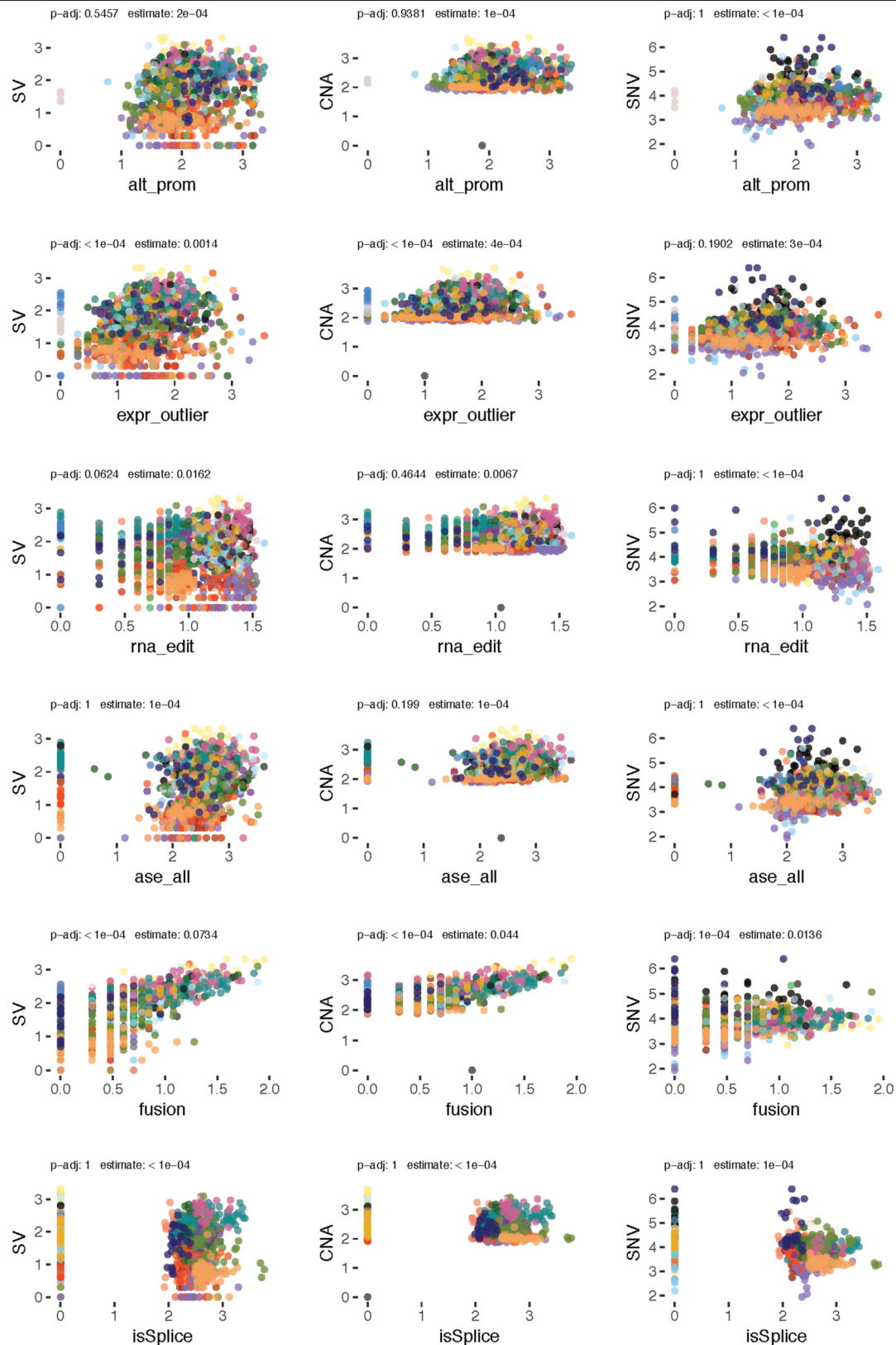
**Extended Data Fig. 18 | Recurrent and promiscuous RNA fusions. a**, Features of the 27 most recurrent in-frame or open-reading-frame-retaining fusions. Kinase column indicates whether one of the gene partners is a kinase gene **b**, Network with connected clusters of at least 10 genes. Genes are represented as nodes, and the size of a node is proportional to the number of gene-fusion partners. Two nodes are connected if one fusion was detected involving the

two genes: an edge is coloured blue if the fusion has evidence for matched structural rearrangements and is coloured red otherwise. Nodes and connections are shown only between promiscuous genes. The colour intensity indicates whether a gene is involved more often in a fusion as a 3' (purple) or 5' (green) gene or both (white).



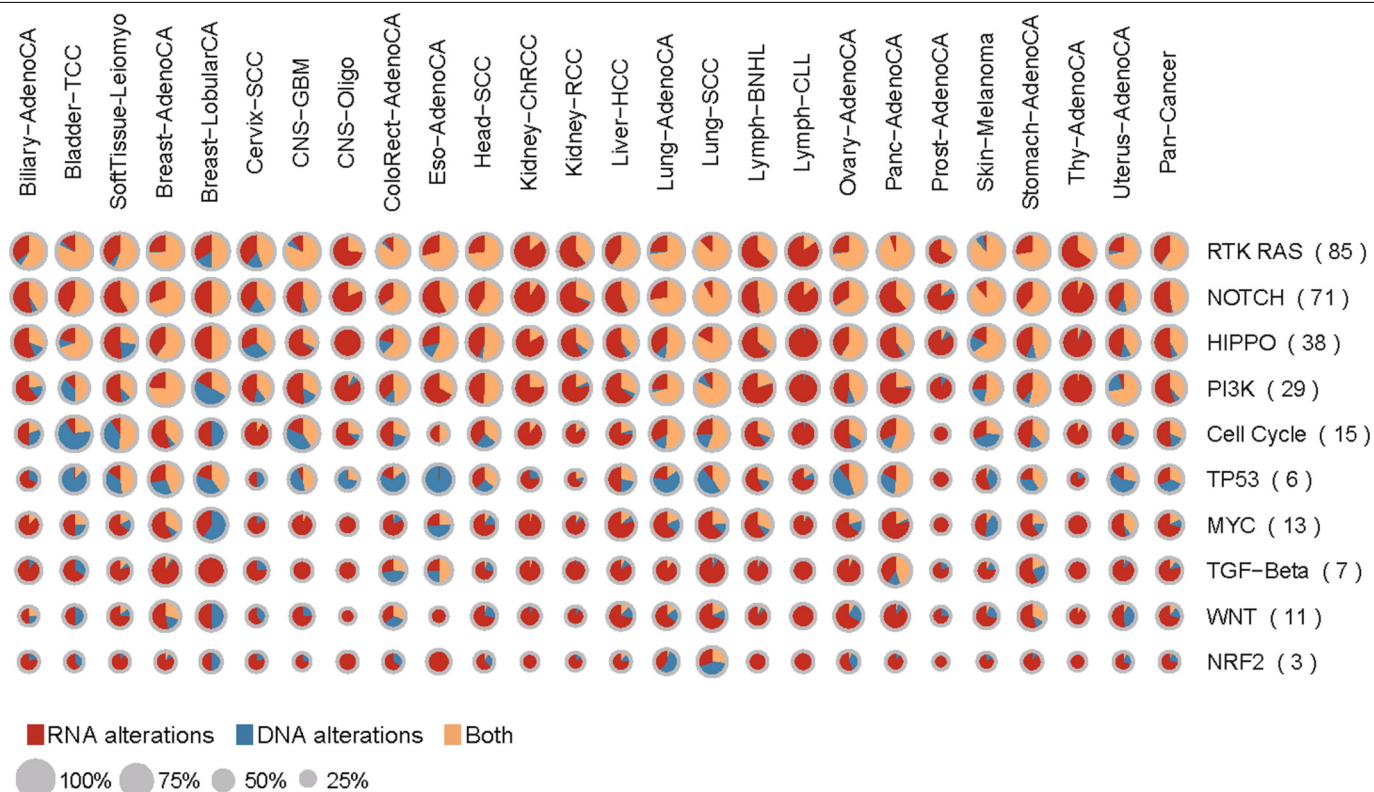
**Extended Data Fig. 19 | Structural rearrangements associated with RNA fusions.** **a**, Systematic classification scheme of all gene fusions based on underlying structural variants (SVs). Numbers of fusion events of different classes are shown to the right. **b**, Schematic of examples of different types of structural-variant-supported fusions: (1) direct fusions; (2) intercomposite fusions; and (3) intracomposite fusions. Bridged fusions are shown in Fig. 3b. Only one of the possible orders of genomic arrangement is depicted in each case, with break points highlighted by thunderbolts. **c**, Supported rearrangements for composite fusions bring the fused segments of two genes

significantly closer. Natural distance indicates the native distance between two related structural variant break points. Effective distance indicates the distance between the final two break points of the intra- and intercomposite fusions. **d**, The break points of structural-variant-independent fusions are typically closer than those for other interchromosomal fusions, which indicates that at least some of the structural-variant-independent fusions may occur directly at the RNA level, mediated either by *trans*-splicing or read-through events.



**Extended Data Fig. 20 | Correlation of the number of somatic genomic alterations with RNA alterations.** Scatter plots of log<sub>10</sub>-transformed frequency of DNA alterations versus log<sub>10</sub>-transformed frequency of RNA alterations, in which each row is a DNA alteration in the following order: structural variants, copy-number aberrations and non-synonymous variants. Each row is an RNA alteration in the following order: expression outliers, RNA

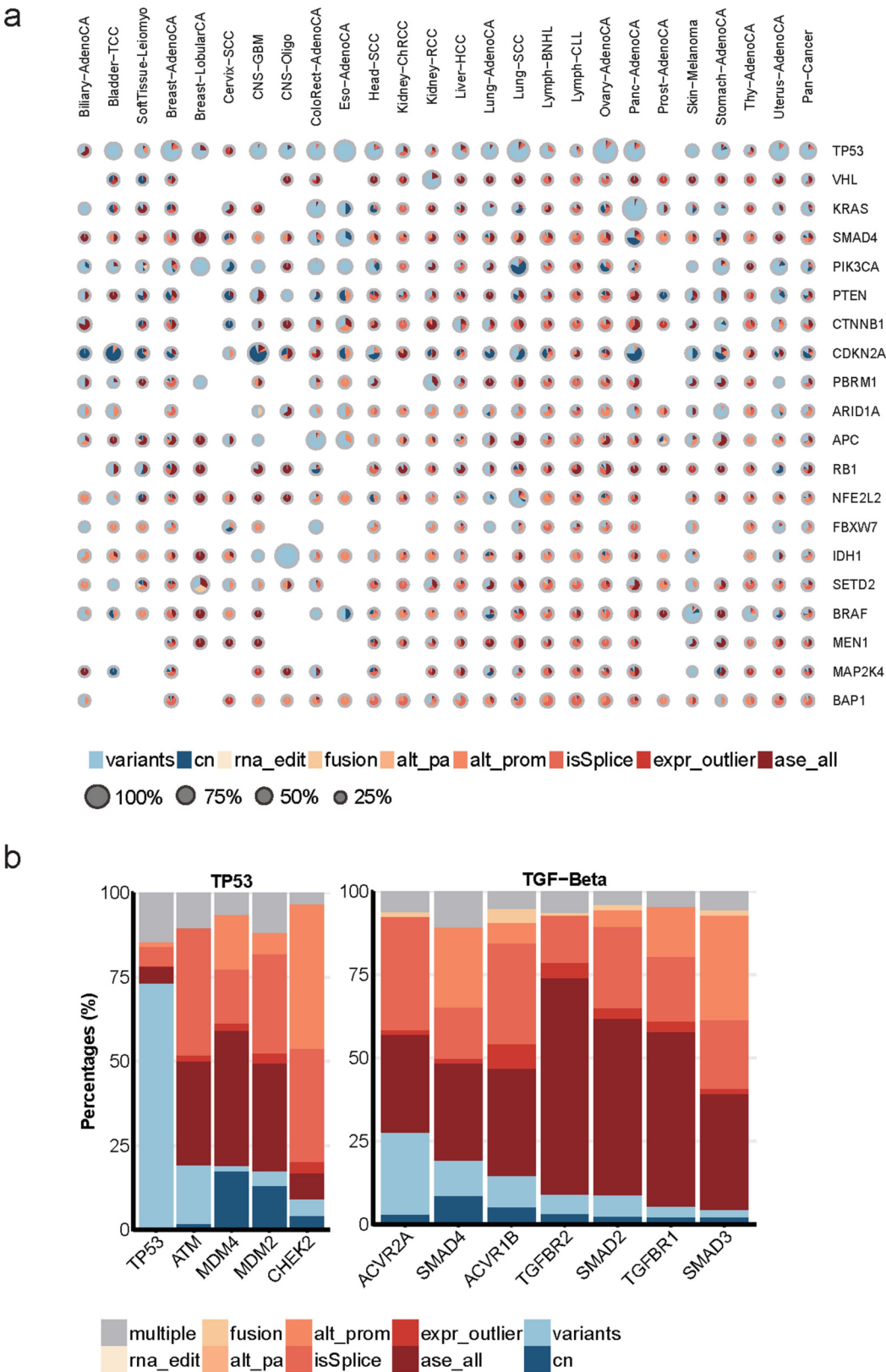
editing, ASE, fusions and splicing. Each point is a sample coloured by histotype, and its position is the log-transformed number of aberrations found in each sample. The Benjamini-Hochberg-adjusted *P* values are calculated from a likelihood ratio test assuming negative binomial distribution; histotype is used as a confounder.



**Extended Data Fig. 21 | Global view of DNA and RNA alterations affecting cancer pathways.** Composite pie charts showing the percentages of RNA alterations, DNA alterations or both, affecting sets of genes in well-characterized cancer pathways and known to be functionally altered in cancer.

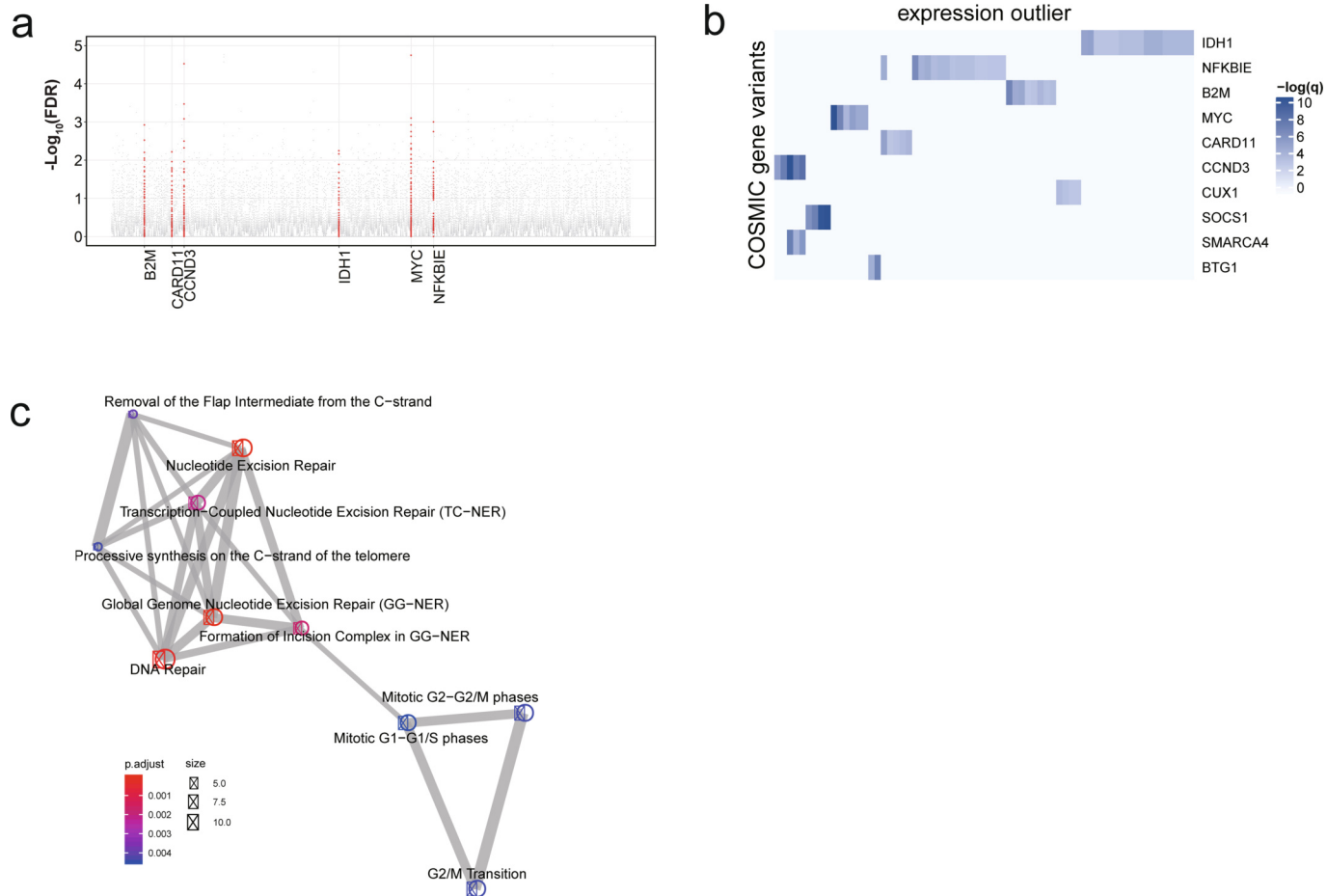
The sizes of circles represent the percentages of patients affected based on the given gene set. The columns indicate different types of cancer. The numbers in parenthesis indicate the number of genes analysed for the specific pathway.





**Extended Data Fig. 22 | Breakdown of DNA and RNA alterations of cancer genes.** **a**, Composite pie charts showing percentages of DNA and RNA alterations for top cancer-driver genes. The 20 most significant cancer-driver genes identified by the PCAWG group in pan-cancer level are depicted, with the sizes of the pie charts indicating the percentages of patients carrying alterations in the given driver gene. The areas represent the relative

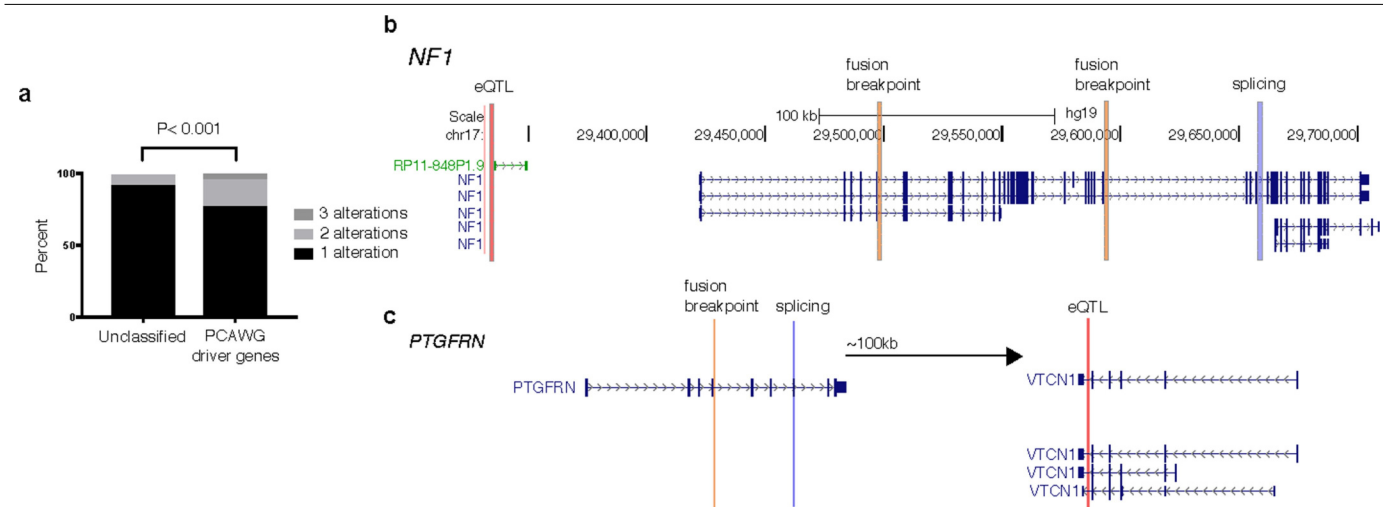
percentages of patients exhibiting different alterations depicted by corresponding colours. When several types of alteration in one pathway affect the same patient, only a fraction is counted towards each type of alteration. **b**, Proportional bar plots showing the distribution of gene alterations for genes in the *TP53* and *TGFB* pathways.



**Extended Data Fig. 23 | *Trans*-associations found by co-occurrence analyses.**

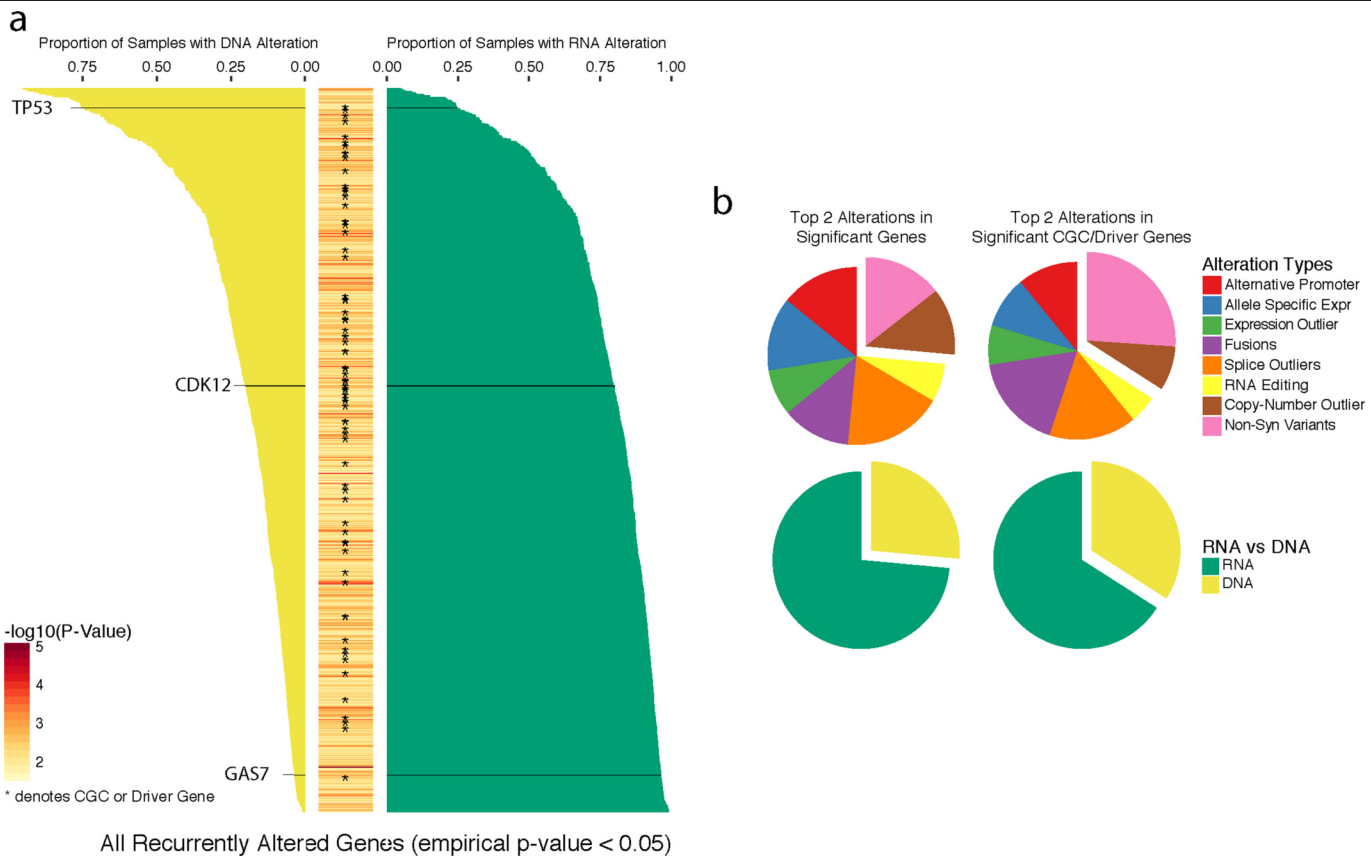
**a**, Scatter plot for association of gene expression outliers with cancer gene variants. Each dot represents an alteration pair. The x axis shows all COSMIC genes ordered alphabetically and the y axis represents the FDR-adjusted  $P$  values ( $q$  values) based on Fisher's exact tests. COSMIC genes with more than five significant associations ( $FDR < 5\%$ ) are coloured in red and labelled. **b**, Heat map showing the extent of associations between COSMIC gene somatic mutations and expression outliers of all genes. Each row indicates one gene,

and the colour intensity shows the significance of *trans*-association. COSMIC genes labelled to the right are ordered by the number of significant associations. Only the top 10 genes are shown. **c**, Enrichment map showing the significant ( $FDR \leq 0.01$ ) pathways based on the top 100 significant genes associated with *B2M* alterations. Colour intensity represents enrichment significance, node sizes the number of analysed genes belonging to the given pathway and edge sizes the degree of overlap between two gene sets. Only the top 10 enriched terms are shown.



**Extended Data Fig. 24 | Genes can be altered in *cis* by several mechanisms.**  
**a**, Genes with at least one type of RNA alteration that also has an associated change at the DNA-level in *cis*. Genes are either classified as a PCAWG driver

gene or not classified as a driver gene or a cancer gene from the cancer gene census. **b, c**, Examples of a known cancer gene, *NF1* (**b**), and an unclassified gene, *PTGFRN* (**c**), having heterogeneous mechanisms of alterations.

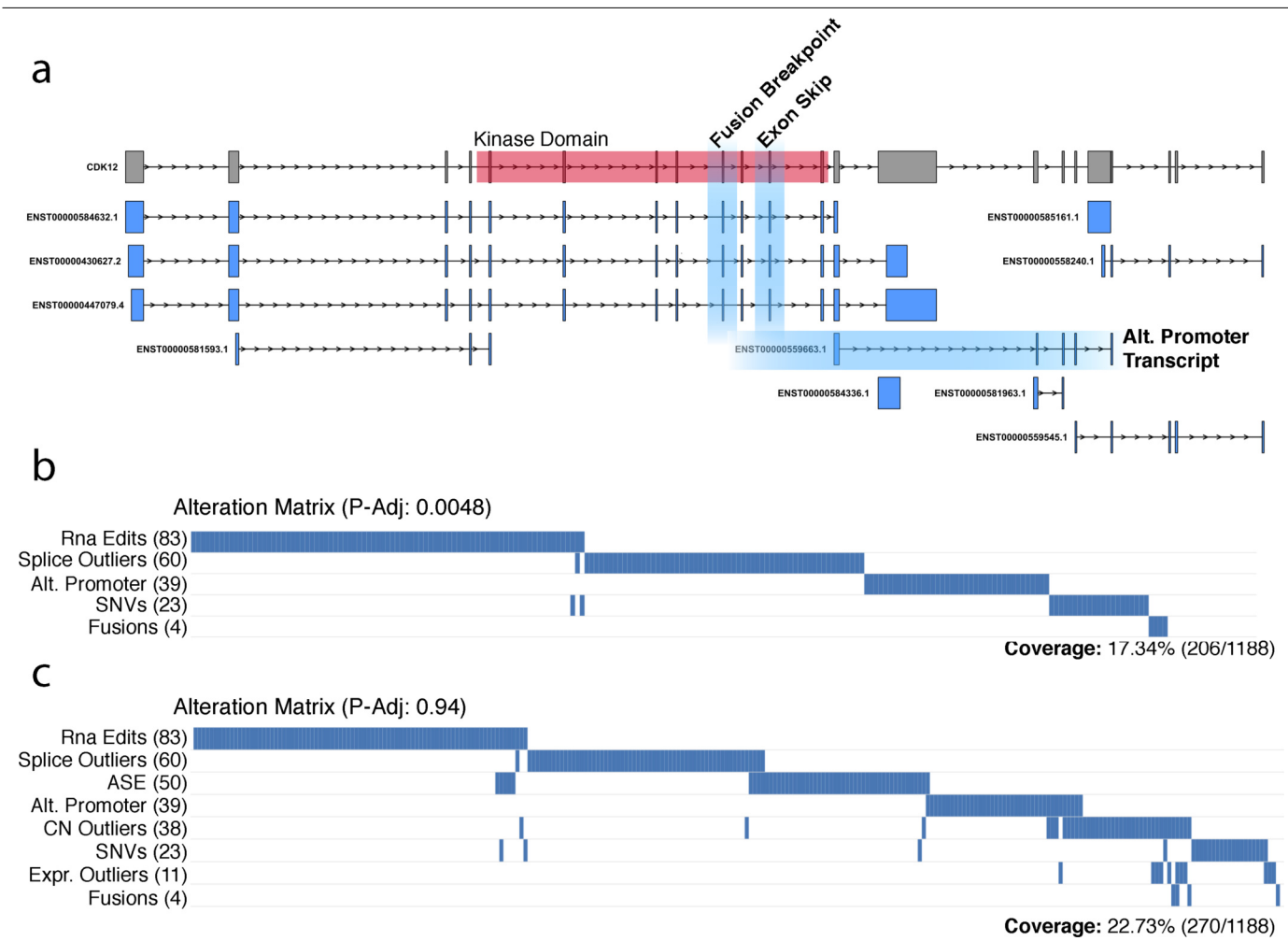


**Extended Data Fig. 25 | Proportion of genes with DNA or RNA alterations.**

**a**, Full list of 731 genes that are both frequently and heterogeneously altered across both RNA- and DNA-level alterations. Yellow bars to the left indicate the proportion of samples that had DNA-level alterations, whereas green bars to the right indicate the proportion of samples with RNA-level alterations. Middle

column is a heat map corresponding to the  $-\log_{10}(P\text{ value})$ . Asterisks indicate a COSMIC Cancer Gene Census (CGC) gene or PCAWG driver genes.

**b**, Distribution of alteration types among all significant genes or just CGC or PCAWG driver genes.



**Extended Data Fig. 26 | Outlier events in *CDK12*.** **a**, Fusion, splicing and alternative promoter outlier events of the RNA alterations that lead to either partial or full removal of the kinase domain in *CDK12*. **b**, All outlier events in *CDK12*, including those not contained directly within the kinase domain, across all 1,188 samples. Each column is a sample and each row is the alteration type.

Although not directly searching for mutually exclusive events across all genes, we find that *CDK12* is marginally mutually exclusive in RNA editing, splicing outliers, alternative promoters, non-synonymous variants and fusions ( $4.810^{-3}$ , unweighted WExT). **c**, All alteration events that occur within *CDK12* across all 1,188 samples, which is not mutually exclusive.



**Extended Data Table 1 | RNA alteration data**

RNA data	Total number of gene alterations found	Mean number of gene alterations per donor
Gene expression (PCAWG)	93,481	78.69
RNA fusions	5,900	4.97
Alternative promoters	246,224	207.26
Alternative splicing	345,115	290.50
Allele-specific expression	544,664	458.47
RNA editing (with a non-synonymous change)	14486	12.19
Combined gene-centric table (DNA and RNA alterations)	1,523,098	1,282.07

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☐ ☒ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

Data collection

Core data was collected through Pan-cancer Analysis of Whole Genomes Data Coordination center through <https://dcc.icgc.org/releases/PCAWG/>

Data analysis

Core RNA-Seq alignment pipelines are available through Github/Docker: [https://github.com/akahles/icgc\\_rnaseq\\_align](https://github.com/akahles/icgc_rnaseq_align), [https://hub.docker.com/r/nunofonseca/irap\\_pcaWG/](https://hub.docker.com/r/nunofonseca/irap_pcaWG/). STAR, TopHat2, HTSeq, Kallisto, Limix, PLINK, Bedtools, Vcftools, Bcftools, Samtools, Tabix, GATK, ASEReadCounter, Lavaan, Mediate, SplAdder, SAVNet, Sv2gf

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium is described here<sup>58</sup> and available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorisation. Data derived specifically from RNA-Seq analysis can be found at <https://dcc.icgc.org/releases/PCAWG/transcriptome>. Subfolders contain identification and quantification of alternative promoter usage, alternative splicing, RNA fusions, gene expression, transcript-level expression, and RNA editing. Identified eQTLs are in <https://dcc.icgc.org/releases/PCAWG/transcriptome/eQTL> and a binarized table indicating all RNA and DNA alterations for each gene can be found in the subfolder [https://dcc.icgc.org/releases/PCAWG/transcriptome/recurrence\\_analyses/](https://dcc.icgc.org/releases/PCAWG/transcriptome/recurrence_analyses/). Additionally, QC metrics and metadata are also included.

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences

### Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study represents the analysis of 1,188 donors from the Pan-Cancer Analysis of Whole Genomes project. The sample size was limited to the availability of RNA-Seq data from matched donors with WGS data from the PCAWG project.
Data exclusions	A larger set of 2,217 RNA-Seq libraries were initially collected and data were excluded after QC analysis. The QC criteria was standard and pre-established before excluding data.
Replication	Reproducibility of the analysis is ensured through data-sharing and code-sharing. Unfortunately, at the time of the analysis there were no appropriate datasets to use for replication studies of associations, since this is one of the largest collections of pan-cancer whole genomes and matched transcriptomes. For the somatic eQTL analysis, there were some related studies that came to similar conclusions and are noted in the Supplementary Information.
Randomization	Cancer histotypes were defined by the PCAWG Pathology and Clinical Correlates Working Group based on tumor histology. These tumor subtypes were accounted for as covariates in all applicable analyses of association.
Blinding	Blinding was not relevant to our study as it was essential to understand underlying confounding variables in our associations, such as tumor subtype, sex, etc.

## Materials & experimental systems

Policy information about [availability of materials](#)

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Research animals
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

# Method-specific reporting

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Magnetic resonance imaging

# The pheromone darcin drives a circuit for innate and reinforced behaviours

<https://doi.org/10.1038/s41586-020-1967-8>

Received: 16 July 2018

Accepted: 12 December 2019

Published online: 29 January 2020

Ebru Demir<sup>1,2</sup>, Kenneth Li<sup>1</sup>, Natasha Bobrowski-Khoury<sup>1,2</sup>, Joshua I. Sanders<sup>2,3</sup>, Robert J. Beynon<sup>4</sup>, Jane L. Hurst<sup>5</sup>, Adam Kepecs<sup>2,6,7\*</sup> & Richard Axel<sup>1,8\*</sup>

Organisms have evolved diverse behavioural strategies that enhance the likelihood of encountering and assessing mates<sup>1</sup>. Many species use pheromones to communicate information about the location, sexual and social status of potential partners<sup>2</sup>. In mice, the major urinary protein darcin—which is present in the urine of males—provides a component of a scent mark that elicits approach by females and drives learning<sup>3,4</sup>. Here we show that darcin elicits a complex and variable behavioural repertoire that consists of attraction, ultrasonic vocalization and urinary scent marking, and also serves as a reinforcer in learning paradigms. We identify a genetically determined circuit—extending from the accessory olfactory bulb to the posterior medial amygdala—that is necessary for all behavioural responses to darcin. Moreover, optical activation of darcin-responsive neurons in the medial amygdala induces both the innate and the conditioned behaviours elicited by the pheromone. These neurons define a topographically segregated population that expresses neuronal nitric oxide synthase. We suggest that this darcin-activated neural circuit integrates pheromonal information with internal state to elicit both variable innate behaviours and reinforced behaviours that may promote mate encounters and mate selection.

Communication through scents elicits innate and learned behavioural repertoires that enhance the reproduction and survival of the species<sup>1</sup>. Male mice deposit scent marks that attract females and enable assessment of the quality and compatibility of potential mates<sup>2,5</sup>. Innate attraction in females is elicited by the non-volatile protein pheromone darcin (MUP20)<sup>3,4</sup>, a member of the major urinary protein (MUP) family that is recognized by receptors in the vomeronasal organ<sup>6</sup>. Darcin not only elicits innate attraction but can also serve as an unconditioned stimulus for both place and odour conditioning, enabling a female to recognize, assess and locate males on the basis of their scent marks<sup>3–5</sup>.

We developed a quantitative behavioural paradigm to examine the effects of darcin, and found that the pheromone elicits a complex and variable behavioural array. Female mice were placed in a chamber equipped with two ports that contained glass fibre filters embedded with different social olfactory cues, and entry to the ports was quantified. The frequency of port entry provides a measure of preference for the cues present on the individual filters. During the initial habituation each port contained a blank filter, and port entries (pokes) were infrequent (mean  $\pm$  s.e.m. poke count: left port  $18 \pm 3$ , right port  $14 \pm 3$ ; Fig. 1b1). The mice were then exposed in their home cage<sup>3</sup> to bedding that had been soiled by male mice, after which the number of pokes increased substantially without any apparent side bias (left port  $247 \pm 35$ , right port  $246 \pm 3$ ; Fig. 1b2). The response to darcin was therefore examined in cycling female mice after exposure

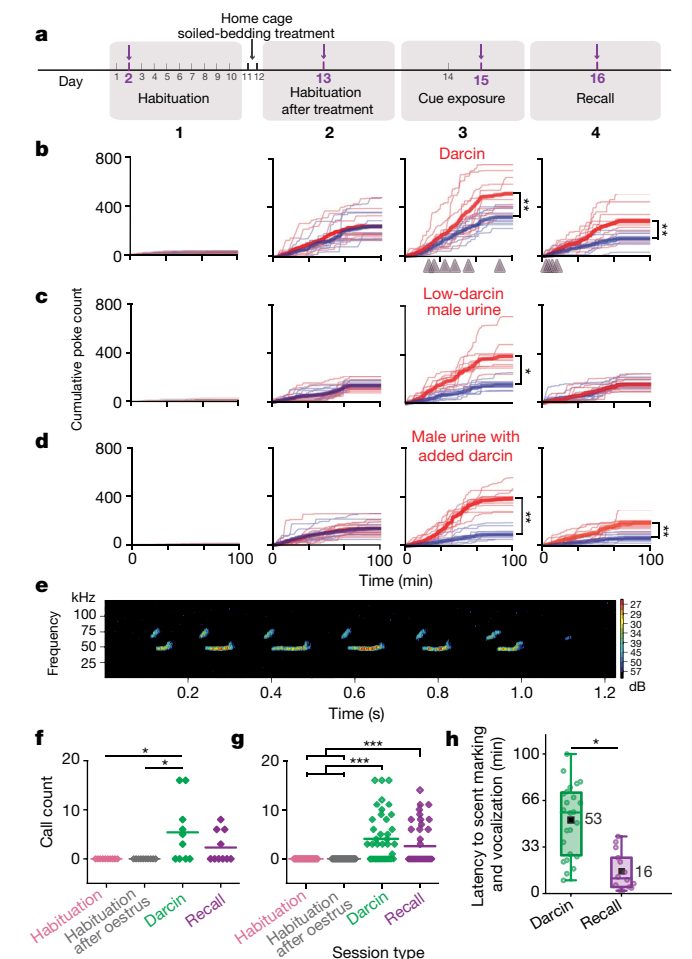
to male-soiled bedding<sup>3</sup>. Poke frequency was higher for the port that contained the recombinant darcin (darcin-containing port  $516 \pm 47$ , blank  $326 \pm 21$ ; Fig. 1b3). Exposure to urine that contained very low levels of darcin (low-darcin urine, from male BALB/c mice)<sup>4</sup> also elicited more frequent port entries than did blank filters in this assay, both with and without the addition of recombinant darcin (low-darcin urine  $386 \pm 42$ , blank  $154 \pm 14$ , Fig. 1c3; recombinant darcin-added urine  $391 \pm 29$ , blank  $96 \pm 18$ , Fig. 1d3).

Innately attractive cues can often serve as a teaching signal, reinforcing both classical and instrumental learning<sup>7</sup>. We examined whether exposure to darcin alone or to low-darcin urine elicits a lasting preference for the darcin port after the stimulus is removed. Female mice were exposed to a social cue in one port and then placed into a clean chamber on the following day with blank filters in both ports. Poke counts were significantly greater in the port that had previously contained either darcin ( $285 \pm 38$ , blank  $146 \pm 16$ ; Fig. 1b4) or urine with equivalent levels of darcin ( $179 \pm 15$ , blank  $65 \pm 9$ ; Fig. 1d4). By contrast, exposure to urine with very low levels of darcin did not result in a port preference during recall sessions on the following day (prior exposure to low-darcin urine  $147 \pm 14$ , blank  $147 \pm 16$ ; Fig. 1c4). Therefore, both low-darcin male urine and darcin elicit a port preference, but only exposure to normal levels of darcin results in a remembered preference.

We also observed that female mice that were exposed to darcin emitted ultrasonic vocalizations and engaged in urinary scent marking

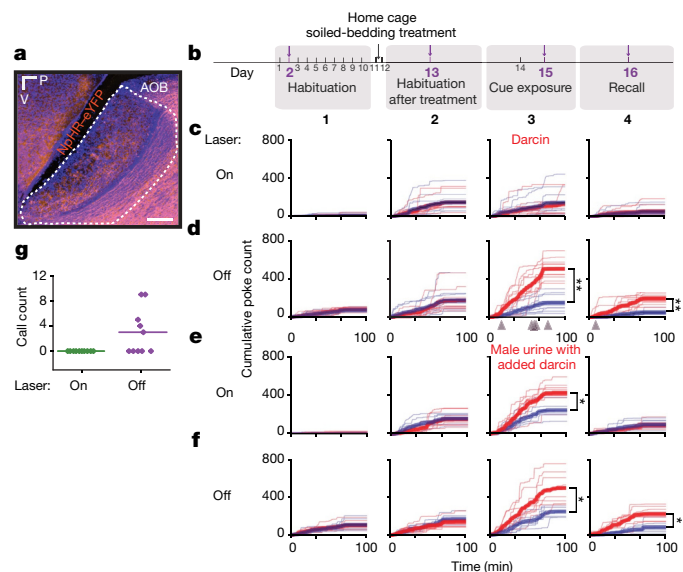
<sup>1</sup>Department of Neuroscience, Columbia University, Mortimer B. Zuckerman Mind Brain Behavior Institute, New York, NY, USA. <sup>2</sup>Department of Neuroscience, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. <sup>3</sup>Sanworks LLC, Stony Brook, NY, USA. <sup>4</sup>Centre for Proteome Research, Institute of Integrative Biology, University of Liverpool, Liverpool, UK. <sup>5</sup>Mammalian Behaviour and Evolution Group, Institute of Integrative Biology, University of Liverpool, Leahurst Campus, Neston, UK. <sup>6</sup>Department of Neuroscience, Washington University School of Medicine, St. Louis, MO, USA. <sup>7</sup>Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. <sup>8</sup>Howard Hughes Medical Institute, Columbia University, New York, NY, USA. \*e-mail: akepecs@wustl.edu; ra27@columbia.edu





**Fig. 1 | Darcin elicits an array of behaviours.** **a**, Timeline of the two-port preference assay. **b–d**, Cumulative poke counts in ports containing darcin ( $1 \mu\text{g} \mu\text{g}^{-1}$ ) (**b**), urine from BALB/c male mice with very low darcin levels (less than  $0.1 \mu\text{g} \mu\text{g}^{-1}$ ) (**c**) and BALB/c male urine with added recombinant darcin ( $1 \mu\text{g} \mu\text{g}^{-1}$ ) (**d**) (red), compared with ports containing control filters (blue) during cue-exposure sessions. Counts are shown on days 2, 13, 15 and 16 (graphs 1–4, respectively), as indicated by arrows on the timeline. Mean (bold lines,  $n = 30$  mice) and individual (fine lines) counts are shown. The time-stamps for ultrasonic vocalizations and scent marking are indicated as arrowheads (**b** (3, 4)). Bias in counts was assessed using the two-sided Wilcoxon signed-rank test (**b** (3, 4),  $**P = 0.004$ ,  $n = 10$ ; **c** (3),  $*P = 0.006$ ,  $n = 10$ ; **d** (3, 4),  $**P = 0.004$ ,  $n = 10$ ). **e**, Spectrogram of an example song detected during darcin-exposure sessions. **f**, **g**, Mean call count (horizontal line) and total number of calls made by individual mice (diamonds);  $n = 10$  mice (**f**) as tested in **b**,  $n = 43$  mice (**g**) as tested across the study. Calls were compared using the two-sided Wilcoxon signed-rank test (**f**,  $*\text{adjusted-}P = 0.03$ ; **g**,  $***\text{adjusted-}P = 0.00003$  and  $P = 0.0001$ ). **h**, Latency to urinary marking and vocalization in response to darcin ( $n = 24$  mice) and during recall ( $n = 14$  mice) sessions. Mean (squares) and individual (circles) latencies are shown. The bounds in box plots are defined by the 25th and 75th percentile of the distribution. The line represents the median and the upper and lower whiskers represent 75th percentile +  $1.5 \times$  interquartile range (IQR) and 25th percentile –  $1.5 \times$  IQR, respectively. Latencies are compared using the two-sided Wilcoxon signed-rank test ( $*P = 0.03$ ).

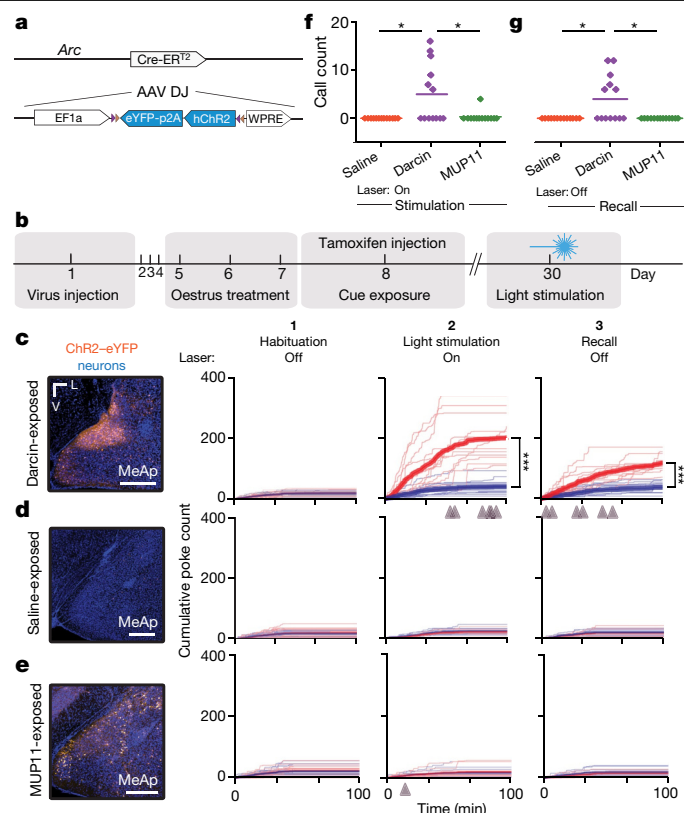
(Fig. 1e–h). Scent marks were located closer to the darcin port (Extended Data Fig. 1a, b) and were smaller in size (Extended Data Fig. 1c) than those observed from free urination, which was consistent with the deliberate deposition of scent close to darcin; this suggests a distinction between the two behaviours. Ultrasonic vocalizations were consistently linked with urinary marking and occurred within 40 ms of one another (mean  $\pm$  s.e.m.  $42 \pm 9$  ms; Supplementary Video 1). These episodes did not occur immediately upon exposure to darcin, but appeared



**Fig. 2 | Optogenetic silencing of the AOB results in the suppression of darcin-evoked behaviours.** **a**, eNpHR-eYFP expression in the AOB. Scale bar,  $200 \mu\text{m}$ . P, posterior; V, ventral; this experiment was independently repeated with 18 mice. **b**, Timeline of the two-port preference assay. **c–f**, Cumulative poke counts with (**c**, **e**) and without (**d**, **f**) optical silencing of the AOB. In **c**, **d** mice were exposed to darcin (3) ( $1 \mu\text{g} \mu\text{g}^{-1}$ ) ( $n = 10$ ) and in **e**, **f** mice were exposed to C57BL/6 male urine (3) with normal levels of darcin ( $1 \mu\text{g} \mu\text{g}^{-1}$ ) ( $n = 8$ ) in one port (red) and a blank filter (blue) in the second port. Mean (bold lines) and individual (fine lines) counts are shown. The time-stamps for ultrasonic vocalizations and scent marking are indicated as arrowheads (**d** (3, 4)). Counts were compared using the two-sided Wilcoxon signed-rank test (**d** (3),  $**P < 0.001$ ; **e** (3),  $f$  (3, 4),  $*P < 0.008$ ). **g**, Mean call count (horizontal lines) and total number of calls made by individual mice (diamonds) during darcin exposure with (**c**, **e**) and without (**d**, **f**) AOB silencing;  $n = 10$  mice. Calls were compared using the two-sided Wilcoxon signed-rank test.

after a long and variable delay during a 100-min session (mean latency  $53 \pm 5$  min, Fig. 1h). Vocalization and urinary scent marking were also observed during recall sessions (Fig. 1f–h). These episodes occurred earlier in the recall session than in the darcin-exposure session (recall sessions  $16 \pm 4$ , darcin sessions  $53 \pm 5$  min; Fig. 1h). Male urine that contained normal levels of darcin also stimulated scent-marking behaviour during cue and recall sessions, but low-darcin urine stimulated marking only when present and not during recall sessions (Extended Data Fig. 2d). Thus, darcin induces a behavioural repertoire that comprises attraction and ultrasonic vocalization simultaneous with urine marking, behaviours that may serve as reciprocal communication. Moreover, this behavioural repertoire is also observed during recall sessions in the absence of darcin.

We next implemented genetic strategies to identify the neural circuitry that mediates these darcin-induced behaviours. Darcin binds to V2R receptors on sensory neurons in the vomeronasal organ<sup>6</sup>. These neurons extend axons through the skull, where they converge to form microglomeruli within the accessory olfactory bulb (AOB)<sup>8</sup>. Microglomeruli are innervated by mitral cells that project to multiple brain regions—including the cortical amygdala, bed nucleus of the stria terminalis, and the medial amygdala (MeA)<sup>8,9</sup>. We demonstrated that this pathway is responsible for the behavioural repertoire elicited by darcin by silencing the AOB. Bilateral injection of an adeno-associated virus (AAV) encoding halorhodopsin<sup>10</sup> fused to enhanced yellow fluorescent protein (eNpHR-eYFP) resulted in the expression of eNpHR-eYFP (Fig. 2a) in the majority of mitral cells in the AOB ( $73 \pm 8\%$  across mice). AOB silencing eliminated the preference for the darcin-containing port ( $180 \pm 49$ , blank  $149 \pm 37$ ; Fig. 2c3) and suppressed darcin-evoked



**Fig. 3 | Activation of darcin-responsive neurons in the MeA recapitulates pheromone-induced behaviours.** **a**, Genetic strategy used to express ChR2 in pheromone-responsive neurons. **b**, Timeline of experimental manipulations. **c–e**, Left, representative images showing eYFP expression in the posterior MeA after exposure to darcin (mean  $\pm$  s.e.m.: eYFP counts, 255  $\pm$  29 in the MeApd and 115  $\pm$  16 in the MeApv) (**c**), saline (16  $\pm$  5 in the MeApd and 23  $\pm$  7 in the MeApv) (**d**) and MUP11 (54  $\pm$  10 in the MeApd and 42  $\pm$  9 in the MeApv) (**e**). Scale bars, 400  $\mu$ m; L, lateral; V, ventral. Right, corresponding cumulative poke counts. Mean (bold lines,  $n$  = 13 mice for each group,  $n$  = 39 mice in total) and individual (fine lines) counts are shown. The time-stamps for ultrasonic vocalizations and scent marking are indicated as arrowheads (**c** (2, 3), **e** (2)). Counts were compared using the two-sided Wilcoxon signed-rank test (**c** (2, 3),  $***P$  = 0.0002). **f, g**, Mean (horizontal lines) and total calls made by individual mice (diamonds) during the light-stimulation sessions (**f**) and subsequent recall sessions (**g**);  $n$  = 13 mice per group. Calls were compared using the two-sided Mann–Whitney test adjusted for multiple comparisons (**f**,  $*P$  < 0.05; **g**,  $*P$  = 0.02).

ultrasonic vocalizations and scent marking (Fig. 2g). By contrast, the preference for male urine with normal levels of darcin was not suppressed during AOB silencing (423  $\pm$  40, blank 243  $\pm$  25; Fig. 2e3). AOB silencing did not affect port investigation during the initial habituation periods with the blank filters (Fig. 2c1, 2–2f1, 2).

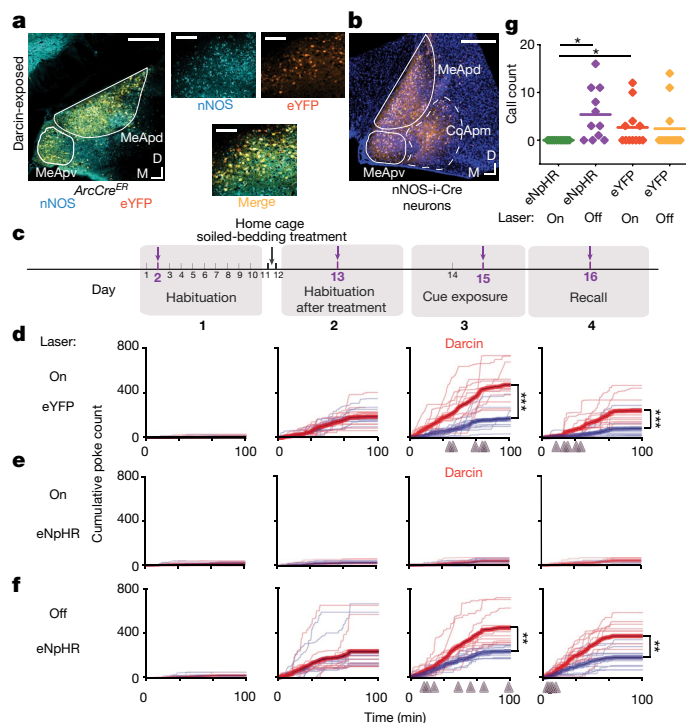
Exposure to recombinant darcin elicited a memory for the darcin port (prior exposure to darcin 190  $\pm$  16, blank 56  $\pm$  6; Fig. 2d4), but a port preference was not observed if the AOB was silenced during darcin exposure (prior exposure to darcin 55  $\pm$  15, blank 60  $\pm$  18; Fig. 2c4). Females that were exposed to male urine containing normal levels of darcin also exhibited a persistent port preference during AOB silencing (urine 423  $\pm$  40, blank 243  $\pm$  25; Fig. 2e3), but failed to show a preference for this port in the recall sessions (prior exposure to urine containing normal levels of darcin 97  $\pm$  16, blank 94  $\pm$  15; Fig. 2e4). These observations show that the AOB is necessary for darcin-induced attraction behaviours as well as for conditioning. Other components of male urine also elicit attraction that is independent of the AOB but fail to reinforce conditioned behaviours.

The mitral cells—the projection neurons of the AOB—send axons to the MeA<sup>8,9</sup>. We identified the neurons of the MeA that are responsive to darcin by using the promoter of the activity-dependent gene *Arc* to express channelrhodopsin-2 (ChR2), a light-gated ion channel<sup>11</sup>. AAV encoding Cre-dependent channelrhodopsin fused to the fluorescent protein eYFP was injected into the MeA of transgenic mice (Arc-CreER mice; Fig. 3a) in which the *Arc* promoter drives the expression of the tamoxifen-sensitive Cre recombinase (Cre-ER)<sup>12</sup>. The administration of tamoxifen followed by exposure to darcin should result in the expression of ChR2–eYFP in neurons that are activated by darcin. We compared the expression of Fos with that of ChR2–eYFP and found that ChR2–eYFP is faithfully expressed in neurons that respond to darcin (78  $\pm$  4% of the ChR2–eYFP<sup>+</sup> neurons also express endogenous Fos, and 79  $\pm$  3% of the neurons that express endogenous Fos also express ChR2–eYFP;  $n$  = 6 mice).

We next determined whether the activation of neurons that express ChR2 after exposure to darcin is sufficient to recapitulate the behaviours elicited by darcin. Arc-CreER mice injected with AAV encoding Cre-dependent ChR2–eYFP in the posterior dorsal medial amygdala (MeApd) and the posterior ventral medial amygdala (MeApv) (Fig. 3a, b) were treated with tamoxifen and then exposed to darcin, saline or a control MUP (MUP11)<sup>3,4</sup>. Histological analysis of ChR2–eYFP expression induced by exposure to darcin revealed a dense clustering of ChR2–eYFP neurons that were restricted largely to the MeApd and the MeApv (Fig. 3c). Exposure to MUP11<sup>3,4</sup> revealed sparser labelling in both the MeApd and the MeApv, and even sparser labelling was observed after exposure to saline (Fig. 3d, e). Mice that expressed ChR2–eYFP induced by exposure to darcin, MUP11 or saline were introduced into the behavioural chamber after two days of habituation. We then photoactivated the MeA when the mice entered one of the two ports containing blank filters, to recapitulate exposure to darcin. Mice that expressed ChR2–eYFP induced by darcin exposure exhibited a strong preference for the stimulation port (mean poke counts: light 202  $\pm$  21, no light 40  $\pm$  7; Fig. 3c2). Photoactivation of the ensemble of darcin-responsive neurons also elicited ultrasonic vocalizations and scent marking (Fig. 3f, Extended Data Fig. 2a–c). Photoactivation of the MeA in mice that expressed ChR2–eYFP after exposure to saline (light 26  $\pm$  3, no light 24  $\pm$  2; Fig. 3d2) or MUP11 (light 19  $\pm$  4, no light 20  $\pm$  5; Fig. 3e2) did not elicit any preferences for the stimulation port, and did not result in ultrasonic vocalizations or urinary scent marking (Fig. 3f).

Mice that expressed ChR2–eYFP in neurons that were responsive to darcin exhibited a remembered preference for the port in which they previously received light stimulation (prior photoactivation 126  $\pm$  8, no activation 39  $\pm$  6; Fig. 3c3). Control mice that expressed ChR2–eYFP in neurons after exposure to MUP11 (prior photoactivation 20  $\pm$  3, no activation 16  $\pm$  2; Fig. 3e3) or to saline (prior photoactivation 20  $\pm$  2, no activation 23  $\pm$  3; Fig. 3d3) exhibited no preference for the previous light-stimulated port. In recall experiments, ultrasonic vocalizations and scent marking were detected only in mice that previously experienced photostimulation of neurons expressing ChR2–eYFP induced by exposure to darcin (Fig. 3g, Extended Data Fig. 2a–c, Supplementary Video 2), and not in mice expressing ChR2–eYFP in neurons activated by exposure to MUP11 or saline (Fig. 3g). We demonstrated that exposure to darcin could also result in conditioned place preference (Extended Data Fig. 3a, b). Therefore, photoactivation of a population of neurons that express ChR2 induced by darcin exposure can elicit innate attraction, ultrasonic vocalizations, urinary scent marking and reinforce conditioned behaviours.

Lactating females fail to exhibit attraction to darcin<sup>13</sup>. We therefore asked whether darcin activates MeA neurons in lactating females. Lactating Arc-CreER mice expressing Cre-dependent eYFP in the MeA were exposed to darcin three to five days postpartum. Exposure to darcin in virgin females resulted in dense labelling of posterior MeA neurons with eYFP expression (mean  $\pm$  s.e.m. eYFP<sup>+</sup> cells: 255  $\pm$  29 in the MeApd and 115  $\pm$  16 in the MeApv). Exposure to darcin during lactation



**Fig. 4 | nNOS neurons in the MeA are necessary for darcin-mediated behaviours.** **a**, Representative image showing the co-expression of eYFP expressed in darcin-responsive neurons (Fig. 3c, left) and nNOS in the posterior MeA (D, dorsal; M, medial;  $n = 7$  mice). Scale bars, 400  $\mu\text{m}$  (main image, left); 100  $\mu\text{m}$  (smaller images, right). **b**, eYFP expression in coronal sections of the posterior MeA of an nNOS-ires-Cre mouse (scale bar, 400  $\mu\text{m}$ ; CoApm, posterior medial cortical amygdala; this experiment was independently repeated with 66 mice). **c**, Timeline of the two-port preference assay. **d–f**, Cumulative poke counts in mice expressing eNpHR (**e, f**) or eYFP (**d**) in nNOS neurons. Neurons expressing eYFP ( $n = 12$  mice) (**d**) and eNpHR ( $n = 11$  mice) (**e**) were photostimulated, with no photostimulation of neurons expressing eNpHR ( $n = 11$  mice) (**f**). Mean (bold lines) and individual (fine lines) counts are shown. The time-stamps for ultrasonic vocalizations and scent marking are indicated as arrowheads (**d** (3, 4), **f** (3, 4)). Counts were compared using the two-sided Wilcoxon signed-rank test (**d** (3, 4),  $***P < 0.0005$ ; **f** (3, 4),  $**P < 0.005$ ). **g**, Vocalization counts of mice expressing eYFP ( $n = 12$  mice) and eNpHR ( $n = 11$  mice). Mean (horizontal lines) and total calls made by individual mice (diamonds). Calls were compared using the two-sided Wilcoxon signed-rank test, adjusted for multiple comparisons ( $*P < 0.05$ ).

resulted in a sparse labelling ( $23 \pm 11$  in the MeApd and  $15 \pm 12$  in the MeApv) at levels similar to that observed upon saline exposure ( $16 \pm 5$  in the MeApd and  $23 \pm 7$  in the MeApv) (Extended Data Fig. 4f–h). By contrast, darcin activates an equivalent number of mitral cells in the AOB of both virgin and lactating females (Extended Data Fig. 4a–e, Fos cells in virgin females  $378 \pm 35$  and in lactating females  $358 \pm 45$ ;  $n = 6$ ,  $P = 0.9$ ). Therefore, the darcin-activated circuit is likely to be gated by lactation in the MeA.

We next identified a genetic marker, neuronal nitric oxide synthase (nNOS), which defines the population of MeA neurons that mediate the darcin-induced behaviours. Immunohistochemical examination of the MeA of Arc-CreER mice revealed that a considerable fraction of neurons that express ChR2-eYFP in response to darcin also express nNOS. We found that 18% of neurons in the posterior MeA express nNOS. Double-labelling experiments demonstrated that this nNOS-expressing population (denoted nNOS neurons) consists of  $55 \pm 4\%$  excitatory neurons (vGlut2<sup>+</sup> cells) and  $24 \pm 3\%$  inhibitory neurons (Gad2<sup>+</sup> cells). We observed that  $74 \pm 2\%$  of the ChR2-eYFP-expressing neurons that are labelled upon darcin exposure also express nNOS, whereas  $66 \pm 3\%$

of the nNOS neurons also express ChR2-eYFP (Fig. 4a). Similar values are obtained in Arc-CreER mice that are exposed to male urine containing normal levels of darcin. The pheromones ESP1<sup>14</sup>, MUP11<sup>3,4</sup> and cat salivary lipocalin Fel-D4<sup>15</sup>, as well as female urine, activated less than 20% of the nNOS neurons (Extended Data Fig. 5, Extended Data Table 1). The majority of the MeA neurons activated by these stimuli do not express nNOS, which demonstrates the specificity of the response of nNOS neurons for darcin.

These observations suggest that activation of the nNOS neurons in the MeA should elicit the behavioural repertoire that is observed upon darcin exposure. We therefore injected AAV encoding Cre-dependent ChR2-eYFP into the posterior MeA of mice in which the *Nos1* promoter drives the expression of Cre (nNOS-ires-Cre) to express channelrhodopsin-2 in nNOS neurons. We then photoactivated nNOS<sup>+</sup> MeA neurons when the mouse entered one of the two ports containing blank filters, and observed a strong preference for the stimulation port (light  $541 \pm 45$ , no light  $66 \pm 12$ ; Extended Data Fig. 3d2). Moreover, photostimulation of nNOS cells expressing ChR2-eYFP evoked ultrasonic vocalization and scent marking (Extended Data Figs. 2a–c, 3f). Photoactivation of these MeA neurons also reinforced conditioned behaviours (prior light  $295 \pm 16$ , no light  $57 \pm 11$ ; Extended Data Fig. 3d3). Control experiments in which AAV encoding Cre-dependent eYFP was injected into the MeA of nNOS-ires-Cre mice failed to elicit any darcin-mediated behaviours upon photostimulation (light  $24 \pm 4$ , no light  $25 \pm 5$ ,  $P = 0.8$ , Extended Data Fig. 3c2; prior light  $23 \pm 7$ , no light  $25 \pm 6$ ,  $P = 0.8$ , Extended Data Fig. 3c3). Therefore, photoactivation of ChR2-eYFP in nNOS neurons in the MeA is sufficient to recapitulate both the innate and the reinforcing behaviours that are observed upon exposure to darcin.

These observations predict that silencing of the nNOS neurons in the MeA should impair the behavioural response to darcin. The bilateral injection of an AAV (AAVDJ-EF1a-DIO.eNpHR3.0-eYFP, Fig. 4b) encoding the Cre-dependent opsin into the nNOS-ires-Cre mice resulted in the expression of halorhodopsin<sup>10</sup> in nNOS neurons in the MeA. In mice in which the nNOS neurons were silenced, no preference was observed for filters that contained recombinant darcin in the poke-preference assay (darcin  $35 \pm 5$ , blank  $39 \pm 5$ ; Fig. 4e3), and darcin elicited no port preference during recall sessions (prior exposure to darcin  $35 \pm 5$ , blank  $42 \pm 6$ ; Fig. 4e4). Ultrasonic vocalizations and urinary scent marking were also eliminated upon light-induced silencing of nNOS neurons (Fig. 4g). As a control we showed that, when photostimulation was terminated, darcin elicited a strong port preference that was also observed during recall sessions (prior exposure to darcin  $375 \pm 40$ , blank  $186 \pm 28$ ; Fig. 4f4). Light-induced silencing in the MeA of mice that expressed eYFP in nNOS neurons failed to inhibit darcin-mediated behaviours (Fig. 4d). Notably, silencing of the MeA also inhibited the port preference that was elicited by urine containing normal levels of darcin (Extended Data Fig. 6b3). These observations suggest that the components in urine other than darcin that elicit port preference also require the MeA. We found that silencing of nNOS neurons resulted in inhibition of poking to control filters after females were exposed to male scent in their home cages (blank  $24 \pm 5$ , blank  $23 \pm 5$ ; Fig. 4e2). We performed additional experiments to demonstrate that the inhibition of darcin-evoked behaviours upon the silencing of nNOS neurons was not due to diminished motivation (Extended Data Fig. 7).

We then asked whether the nNOS neurons in the MeA are also required for the expression of the remembered response. Female mice were exposed to darcin, and then nNOS neurons were silenced only during recall sessions. These mice exhibited a strong preference for the port that had previously contained darcin (prior exposure to darcin  $254 \pm 22$ , blank  $77 \pm 17$ ; Extended Data Fig. 6e4). Darcin-responsive neurons that express nNOS in the MeA are therefore necessary to recapitulate the innate and reinforcement behaviours elicited by darcin. Recall of darcin memory, however, no longer requires this neural population.

The array of properties elicited by darcin suggests that this pheromone does not elicit a simple behavioural response, but rather activates



a complex integrative process that may optimize mate encounters and mate selection. First, the attractive response is rapid and prolonged upon darcin exposure, whereas vocalization and scent marking are variable and often occur after long delays. Activation of the nNOS population of neurons by darcin may therefore elicit a state of 'sexual drive', which increases the probability of individual component behaviours that are suited to enhance the likelihood of mate encounters under different environmental circumstances. Darcin exposure results in exploration and assessment of the darcin source: the urine of a dominant male. In the absence of the male, after active search strategies have failed, the female may emit ultrasonic vocalizations synchronized with scent marking in an attempt to communicate her presence and her current oestrus status to the male.

Second, darcin activation of the nNOS neurons reinforces both contextual and olfactory learning—generic learning processes—which may allow the female to return to the location of the male's scent mark<sup>4</sup> or to track airborne scents of the territorial male<sup>3,5</sup>. The MeA may therefore provide a signal mediated by darcin to midbrain dopamine neurons to reinforce more traditional 'non-social' reinforcement learning<sup>7</sup>. The more stereotypical communication behaviours elicited by darcin—vocalization and scent marking—might also result from reinforcement of a specific set of social behaviours that coordinate a successful mate search. Whereas the nNOS neurons are required for the behavioural and reinforcing effects of darcin, recall of darcin-elicited memory no longer requires this neural population—presumably reflecting the transfer of a learned representation in other brain structures.

Third, we observe that male urine with very low levels of darcin elicits attraction but does not result in reinforcement learning or memory of port preference. This attractive response does not require the AOB but is eliminated upon silencing the nNOS neurons of the MeA. These observations suggest that the MeA is integrating pheromonal information from the vomeronasal pathway with olfactory cues from the main olfactory system to elicit both innate attraction and learning.

Finally, the response to darcin is dependent on internal state. Lactating females fail to exhibit this complex behavioural response to darcin exposure<sup>13</sup>. We found that darcin activates the projection neurons in the AOB in lactating females, but fails to activate the nNOS neurons in the MeA (Extended Data Fig. 4). Taken together, these observations suggest that the nNOS neurons of the MeA integrate internal state with

the pheromonal cues to mediate both innate variable behaviours and reinforced behaviours that may coordinate a successful mate search.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1967-8>.

1. Malte, A. *Sexual Selection* (Princeton Univ. Press, 1994).
2. Wyatt, D. T. *Pheromones and Animal Behavior: Chemical Signals and Signatures* (Cambridge Univ. Press, 2014).
3. Roberts, S. A. et al. Darcin: a male pheromone that stimulates female memory and sexual attraction to an individual male's odour. *BMC Biol.* **8**, 75 (2010).
4. Roberts, S. A., Davidson, A. J., McLean, L., Beynon, R. J. & Hurst, J. L. Pheromonal induction of spatial learning in mice. *Science* **338**, 1462–1465 (2012).
5. Roberts, S. A., Davidson, A. J., Beynon, R. J. & Hurst, J. L. Female attraction to male scent and associative learning: The house mouse as a mammalian model. *Anim. Behav.* **97**, 313–321 (2014).
6. Kaur, A. W. et al. Murine pheromone proteins constitute a context-dependent combinatorial code governing multiple social behaviors. *Cell* **157**, 676–688 (2014).
7. Schultz, W. Neuronal reward and decision signals: from theories to data. *Physiol. Rev.* **95**, 853–951 (2015).
8. Halpern, M. & Martínez-Marcos, A. Structure and function of the vomeronasal system: an update. *Prog. Neurobiol.* **70**, 245–318 (2003).
9. Dulac, C. & Wagner, S. Genetic analysis of brain circuits underlying pheromone signaling. *Annu. Rev. Genet.* **40**, 449–467 (2006).
10. Gradinaru, V., Thompson, K. R. & Deisseroth, K. eNpHR: a *Natronomonas* halorhodopsin enhanced for optogenetic applications. *Brain Cell Biol.* **36**, 129–139 (2008).
11. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).
12. Root, C. M., Denny, C. A., Hen, R. & Axel, R. The participation of cortical amygdala in innate, odour-driven behaviour. *Nature* **515**, 269–273 (2014).
13. Martín-Sánchez, A. et al. From sexual attraction to maternal aggression: when pheromones change their behavioural significance. *Horm. Behav.* **68**, 65–76 (2015).
14. Kimoto, H., Haga, S., Sato, K. & Touhara, K. Sex-specific peptides from exocrine glands stimulate mouse vomeronasal sensory neurons. *Nature* **437**, 898–901 (2005).
15. Papes, F., Logan, D. W. & Stowers, L. The vomeronasal organ mediates interspecies defensive behaviors through detection of protein pheromone homologs. *Cell* **141**, 692–703 (2010).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

# Article

## Methods

### Mice

All surgical and experimental procedures were performed in compliance with the Guide for the Care and Use of Laboratory Animals<sup>16</sup> from the National Institute of Health Standards, and approved by the Cold Spring Harbor Laboratory and Columbia University Medical Center Institutional Animal Care and Use Committees. Experiments were conducted with 279 female mice between 6 and 30 weeks old. Mice were purchased at 4 weeks old and were handled for at least 10 min each day for a minimum of 5 days before experimentation. Surgeries were performed on mice that were 6 weeks old in order to match their brain coordinates to the Allen Reference Atlas. The mouse lines used were as follows: Arc-CreER (a gift from C. Denny at Columbia University; also available from Jackson Laboratory, Jax stock 022357); ICR outbred (CD-1) wild-type mice (Harlan/Envigo); Ai14 (Rosa-CAG-LSL-tdTomato); nNOS-ires-Cre (Jax stock 017526); vGlut-ires-Cre (Jax stock 028863); Gad2-T2a-NLS-mCherry (Jax stock 023140). The nNOS-ires-Cre mice were crossed to ICR outbred mice (Harlan/Envigo) for 15 generations to exchange their genetic background to the ICR mice. Throughout the study, five mice were co-housed in a single cage for two to six months. This long-term co-housing has the potential to suppress oestrus cycling in females (the Lee–Boot effect)<sup>17</sup>. To ensure that all females had previously encountered male scent and were showing normal oestrus cycling, females were exposed to male-soiled bedding from an unfamiliar strain for at least 60 h<sup>3</sup>. They were then visually evaluated for their stage of oestrus before the experimental testing. One hour before testing, each mouse had its vaginal opening photographed for evaluation. After oestrus entrainment, most females (more than 90%) were evaluated to be in the pro-oestrus stage<sup>3</sup> of the cycle (with swollen, moist, pink and wide-open vaginal openings<sup>18</sup>) and advanced into behavioural testing. Mice were kept in a controlled 12-h day/night (7:00 to 19:00) cycle and tested only during the night phase (23:00 to 6:00).

### Behavioural assays

Before behavioural training, mice were handled for 10 min each day for five days, and were given access to a mouse exercise cage that was enriched with spinning discs and toys for one hour every day during the experimental period. Training took place in a custom-designed sound-isolation chamber containing a behavioural arena (25 × 25 × 28 cm) integrated with two stimulus ports (circular nose port (4.6-cm diameter) with an attachable circular cup for the filter (1.3-cm diameter)), which were surrounded by distinct visual stimuli (stripe and circle stickers were used on either side (Context Kit for Conditioned Place Preference, Stoelting)) on the walls. Mice were tested under room light during the night phase of their day/night cycle (23:00 to 6:00). Mice poked their snouts into stimulus ports to sample the social stimuli. The social cue was presented on a glass microfibre filter in a portable cup attached to the nose port. Social cue ports were constructed out of metal and boiled in detergent (1–2% Alconox for at least 15 min), rinsed thoroughly with water, dipped in 3% hydrogen peroxide and ethanol, rinsed again with running distilled water and air-dried to clean off any contaminants between experiments. The frequency and duration of nose pokes were quantified by means of an infrared beam within the port. The behavioural nose poke data were acquired through a MATLAB interface and a Bpod.

Ultrasonic vocalizations in the chamber were captured using an Avisoft ultrasound microphone with a frequency range of 20–200 kHz. The microphone was connected to a portable time-code generator and reader (Horita PTG2), which generated a time code that was embeddable into both the audio and the video files. Avisoft Recorder USGH software was used to record vocalizations and integrate time codes from the PTG2. To capture urinary scent-marking behaviours with the embedded time code, a Marshall Genlock 3G-SDI HDMI camera was mounted at the base of the transparent chamber. An AJA K; Pro

Recorder, which was connected to the camera and the PTG2, was used to record video for the entire duration of the session. The time code generated by the PTG2 was visible as a display within the video window of the Marshall camera recording through the AJA recorder, and was also recorded by Adobe Captivate.

The nature of the ultrasonic vocalizations in each session was analysed with Avisoft SAS Lab Pro (Supplementary Videos 1, 2). We quantified call counts as the number of syllables in a given session of an individual mouse. Comparison of the calls emitted in response to the pheromone and the calls emitted upon the photoactivation of MeA neurons confirmed that the pheromone- and photoactivation-evoked syllables shared similar sonic qualities (Extended Data Fig. 8, Extended Data Table 2). All spectrograms were additionally parametrized using SAP 2011<sup>19</sup> and MUPET<sup>20</sup> software, and all syllables emitted by the mice during the sessions were manually extracted and classified for analysis (Extended Data Fig. 8). To analyse the urinary scent-marking behaviour of the mice, Adobe Premiere Pro was used. To determine the concurrency between urination and vocalization, Adobe Premiere Pro was used to align the video to the audio by using the time shown by the OLED display of the PTG2 (visible in the video window) in conjunction with the time code encoded in the audio file as temporal references. In addition, engagement of the poke port resulted in the simultaneous activation of a red LED, which was visible to the human eye in the video window but not to the mice, and a TTL (transistor–transistor logic) pulse, which was recorded in the ultrasonic audio track as a labelled time event by the Avisoft Recorder USGH software. Engagement of the port was thus used as an additional online reference to observe the alignment of audiovisual events, and this was recorded by Adobe Captivate. In addition, the distances from urinary drops to the base of each of the ports were quantified for the pheromone, photoactivation and free-urination sessions. Distances were extrapolated from individual frames of the video using Adobe Photoshop.

Mice were placed in the behaviour chamber for 100 min once per day for each session during the dark phase (23:00 to 6:00) of their day/night cycle (7:00 to 19:00). The behavioural chamber and the stimulus ports were thoroughly cleaned with 1–2% Alconox detergent, distilled water, 3% hydrogen peroxide and 80% ethanol, rinsed again with distilled water and air-dried in between individual sessions. The first ten sessions served as habituation sessions, during which no social cue was present in either social cue port. Therefore, there were no special cues available to the mice as they acclimatized to movement in the chamber and, for subjects involved in optogenetic experiments, movement while tethered to the patch cord. For behavioural testing, all mice—except for the optically activated mice—were exposed for 60 h in their home cage<sup>3</sup> to bedding soiled by male mice, followed by an extra habituation session with blank filters in both stimulus ports after this home-cage treatment. Subsequently all mice were tested with social cues or optical activation present in either port. The social cue or activation sides were randomly assigned between two ports across mice to control for any potential side bias. For the optical activation experiments, a nose poke into the stimulation port triggered an external laser pulse (473-nm light, 60 pulses, 20 Hz) using a PulsePal<sup>21</sup> device.

The ICR background mice that were not tested optogenetically were subjected to the following social cues in one port: recombinant darcin (1 µg µl<sup>-1</sup>); male urine with low levels of darcin (<0.1 µg µl<sup>-1</sup> in BALB/c J Ola-Hd urine, purchased from Harlan/Envigo)<sup>4</sup>; male urine with normal adult levels of darcin (1 µg µl<sup>-1</sup>, C57BL/6J Ola-Hsd urine, purchased from Harlan/Envigo)<sup>4</sup>; or recombinant darcin added to BALB/c J Ola-Hd male urine with low levels of darcin (BALB/c J plus recombinant darcin, 1 µg µl<sup>-1</sup>). In all instances, there was no odour in the other port. To confirm the presence or absence of darcin (18,893 Da MUP20), 12% SDS–PAGE gel electrophoresis of all urine samples was performed<sup>3,4</sup>.

The C57BL/6 Arc-CreER and nNOS-ires-Cre/ICR mice tested with optical activation were subjected to optical activation in one port and no optical activation in the other port. The ICR outbred mice tested with



AOB inactivation were subjected to either recombinant darcin (11 µg in 10 µl) or male urine with normal levels of darcin (10 µl of C57BL/6 Ola-Hd urine)<sup>4</sup> in one port and no odour in the other port. The nNOS-ires-Cre mice tested with MeA inactivation were subjected to darcin or male urine with normal levels of darcin (C57BL/6 Ola-Hd)<sup>4</sup> in one port and no odour in the other port. All the optical-silencing experiments used a continuous light-on protocol during the entire test sessions. The final session for all mice was a recall session designed to quantify retained poke preferences in the absence of social cues or optical activation.

For conditioned place preference experiments, C57BL/6 Arc-CreER and nNOS-ires-Cre/ICR mice were introduced into a two-chamber conditioned place preference arena (22 × 16 × 28 cm, length × width × height) for 100 min once per day for each session. The two chambers had distinct walls decorated with visual cues (stripes and circles stickers, Context Kit for Conditioned Place Preference, Stoelting)); chambers were separated by a corridor and a divider, each containing a single nose port. Light stimulation was delivered to a port in one of the two chambers, and there was no optical activation in the other port. A nose poke into the light-stimulation port triggered an external laser pulse (473-nm light, 60 pulses, 20 Hz) using a PulsePal<sup>21</sup> device during the light-stimulation sessions only. During the habituation and recall sessions, a nose poke into the light-stimulation port did not trigger a laser pulse. Videos were recorded throughout the 100-min sessions. The positions of the mice were tracked using Ethovision, and the occupancy trajectories and time spent in each chamber were computed for analysis.

To demonstrate that MeA nNOS neurons are indispensable for social cue reinforcement behaviours only, we optogenetically silenced nNOS neurons in the MeA and tested the mice with water as a reinforcer rather than darcin. Mice were tested using a two-port setup without any social cues. Before behavioural training, mice were gradually water-restricted over the course of a week and kept under water scheduling until the tests were concluded. Mice were placed in the behaviour chamber for 100 min once per day for each session during the dark phase (23:00 to 6:00) of their day/night cycle (7:00 to 19:00). The first ten sessions served as habituation sessions, during which no cue was present in either port. Mice were acclimatized to the movement in the chamber while being tethered to the patch cord. They were then subjected to male-soiled bedding exposure for 60 h in their home cage and an extra habituation session with blank filters in both ports following this home-cage treatment. Subsequently, all mice were tested for cue sessions. The cue sides were evenly split in a random manner between two ports across the mice to control for any potential side bias. During cue sessions, a nose poke in one port rewarded the mice with 5 µl of water, and there was no reward for a nose poke in the other port. Behavioural training sessions lasted 100 min, during which the mice typically collected at least 4 ml of water. The final session for water-reinforcement behaviour was a recall session designed to quantify the retained poke preferences without any water reward. The behavioural hardware was controlled by custom MATLAB programs and a Bpod and PulsePal<sup>21</sup>.

Investigators were blinded to the allocation of the mice during the experiments and data analysis.

### Stereotactic surgeries

An adeno-associated virus (AAV) DJ serotype<sup>22</sup> ( $1.3 \times 10^{13}$  vg ml<sup>-1</sup> (genomic),  $8 \times 10^8$  IU ml<sup>-1</sup> (infectious) titre, Stanford Vector Core Facility) carrying EF1a DIO hChR (E123T/T159C)-p2A-eYFP-WPRE, EF1a DIO NpHR3.0-eYFP, EF1a DIO-eYFP or EF1a NpHR3.0-eYFP construct was injected in 4- to 6-week-old mice. The mice were anaesthetized with an intraperitoneal injection of a ketamine and xylazine mixture (0.13 mg per g body weight ketamine and 0.01 mg per gram body weight xylazine). Small craniotomies were made above the posterior MeA (-2.0 mm AP and 2.3 mm ML from the bregma) or the AOB (3.2 mm AP, 1 mm ML and 0.8–1.5 mm DV). Virus was injected with a glass micropipette using a Picospritzer (General Valve). For posterior MeA injections, 20–60 pulses of 10-ms duration were delivered at 0.2 Hz starting from

a depth of 4.6 mm from the brain surface up to 5.2 mm in 200-µm steps, waiting a minimum of 10 min per site to allow diffusion of the virus. After virus injection, fibre-optic cannulas were implanted. The mice received a supplementary dose of ketamine at 30- to 90-min intervals to maintain the depth of anaesthesia. The cannula was positioned with the help of a stereotaxic arm (David Kopf Instruments) and cannula holder (Doric Lenses) above the craniotomy. The optical cannula was gradually lowered close to the viral injection depth (100–300 µm above the injection site). Two miniature watch screws (Micro-Mark) were fixed into the parietal plates as anchors. The cannula was secured to the skull with light-curable dental cement (Vitrebond Plus) followed by a layer of black dental acrylic (Lang Dental Manufacturing). For post-operative analgesia, ketoprofen (5 mg per kg body weight) was administered subcutaneously. The mice were allowed to recover for one week.

### Exposure of Arc-CreER mice to social cues

One week after stereotaxic viral infection and cannula surgery, 6- to 8-week-old Arc-CreER mice were transferred to a reverse day/night cycle. They were individually housed unless mentioned otherwise and oestrus was synchronized through exposure to male-soiled bedding for 60 h<sup>3,4</sup>. Mice were then injected with 2 mg of tamoxifen (Sigma T5648), which was prepared as a 10 mg ml<sup>-1</sup> stock solution dissolved in a mixture of ethanol and sunflower seed oil (Sigma S5007). Five hours after tamoxifen injection, the mice were exposed to darcin, MUP11, saline, cat salivary lipocalin (Fel-D4)<sup>15</sup>, ESP1<sup>14</sup> (exocrine-gland secreting peptide), male urine with normal levels of darcin, female urine or male urine with low levels of darcin on a glass microfibre filter (10 mm diameter) placed through the roof of their home cage; 10 µl (equivalent to 11 µg of darcin<sup>3,4</sup>, MUP11<sup>3,4</sup>, equivalent to 3.3 µg of Fel-D4<sup>15</sup> and 25 µg of ESP1<sup>14</sup>) was used. The lactating females were separated from their pups five hours before tamoxifen injection and exposed to recombinant darcin between postpartum days 3 and 5. Recombinant cat Fel-D4 was produced using the pMAL Protein Fusion and Purification System (New England Biolabs) and assayed by SDS-PAGE. The mouse ESP1 was synthesized by Atlantic Peptides. The mice were monitored with infrared cameras to confirm that they had interacted with the filters. Optical activation experiments were conducted three weeks after exposure to the cues.

Three weeks after the tamoxifen injection, the Arc-CreER mice that were subjected to optical stimulation were re-exposed to darcin for 2 h and then euthanized for immunohistochemistry.

### Immunohistochemistry

Once the behavioural criteria for each behaviour assay were met, the mice were anaesthetized with a ketamine and xylazine mixture (0.30 mg per g body weight ketamine, 0.03 mg per g body weight xylazine) and perfused transcardially with 4% paraformaldehyde (PFA) in a phosphate buffer pH 7.4 (PBS). The brain was dissected and incubated at 4 °C in 4% PFA, washed in 1× PBS, and stored in PBS at 4 °C until sectioning. Subsequently, 50-µm coronal brain sections were made using a Leica VT1000S vibratome. The sections were incubated with a blocking solution (5% normal goat serum and 0.1% Triton in PBS (PBST)), washed in 0.1% PBST (3 washes, 15 min each) and incubated overnight at 4 °C with primary antibodies diluted in blocking solution. The following primary antibodies were used: anti-GFP (rabbit polyclonal, 1:1,000, Rockland), anti-GFP (chicken polyclonal 1:400, Aves Labs), anti-nNOS (rabbit polyclonal, 1:400, Invitrogen), anti-mCherry (rat monoclonal, 1:800, Thermo Fisher Scientific) and anti-Fos (goat and rabbit polyclonal, 1:500, Santa Cruz Biotechnology; guinea pig polyclonal, 1:5,000, with RRID: AB\_2814707, generated by S. Brenner-Morton, at ZMBBI, Columbia University). The following day the sections were washed in 0.1% PBST (3 washes, 15 min each) and incubated for 2 h at room temperature with secondary antibodies at 1:500 dilutions (Alexa-594 goat anti-rabbit, Alexa-633 donkey anti-goat, Alexa-488 goat anti-rabbit, Alexa-488 goat anti-chicken, Alexa-594 goat anti-rat, Alexa-488 goat anti-guinea pig, Jackson ImmunoResearch, and NeuroTrace Alexa-640/660, Molecular

# Article

Probes). Sections were washed in 1× PBS for 15 min and mounted using Vectashield mounting medium (Vector Laboratories). Confocal images were acquired using an LSM780 Zeiss microscope at 10×, 20× and 65× magnifications. Area and cell counts were manually quantified using ImageJ (NIH) software by an individual who was blinded to the experimental conditions.

## Statistical analysis

Port preferences within each session type (habituation day 2, habituation day 13, cue exposure and recall) for each subject were compared by matched Wilcoxon signed-rank tests. Port bias for the left port over the right port was computed by taking the difference in total poke count between the left and the right ports for each mouse. Comparisons were across each session (habituation day 2, habituation day 13, cue exposure, and recall) using the Wilcoxon signed-rank test. Port bias was compared across independent treatment cohorts by Mann–Whitney (for pairwise comparisons) and Kruskal–Wallis tests (for three-way comparisons). All poke count data did not approximate to normality and so non-parametric tests were used. Call counts were compared across independent cohorts of mice using a Mann–Whitney test and across different sessions of the same cohort of mice using the Wilcoxon signed-rank test. Adjusted *P* values were reported where multiple comparisons were made on the same sample set by using the Holm’s sequential Bonferroni correction method. The probabilities of urinary scent-marking behaviour were compared across sessions using the McNemar test. Exact tests were performed for all comparisons, including those in which the sample sizes were small (the discordant pairs in some of our comparisons were less than 25). The mean latencies to first urinary scent marking were compared using a paired *t*-test. The latency data approximated to normality as confirmed by Shapiro–Wilk, Lilliefors, Kolmogorov–Smirnov, Anderson Darling, D’Agostino–K squared and Chen–Shapiro tests. All analyses were performed using R, OriginLab and MATLAB.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The datasets generated and/or analysed during the current study are available from the corresponding authors on reasonable request.

16. National Research Council (US) Committee. *Guide for the Care and Use of Laboratory Animals* (National Academies Press, 2011).
17. Champlin, A. K. Suppression of oestrus in grouped mice: the effects of various densities and the possible nature of the stimulus. *J. Reprod. Fertil.* **27**, 233–241 (1971).
18. Byers, S. L., Wiles, M. V., Dunn, S. L. & Taft, R. A. Mouse estrous cycle identification tool and images. *PLoS ONE* **7**, e35538 (2012).
19. Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Mitra, P. P. A procedure for an automated measurement of song similarity. *Anim. Behav.* **59**, 1167–1176 (2000).
20. Van Segbroeck, M., Knoll, A. T., Levitt, P. & Narayanan, S. MUPET—Mouse Ultrasonic Profile Extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations. *Neuron* **94**, 465–485.e5 (2017).
21. Sanders, J. I. & Kepecs, A. A low-cost programmable pulse generator for physiology and behavior. *Front. Neuroeng.* **7**, 43 (2014).
22. Grimm, D. et al. In vitro and in vivo gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82**, 5887–5911 (2008).

**Acknowledgements** We thank Z. J. Huang, R. Paik, H. Taniguchi, G. Enikopolov, S. P. Ranade and D. Kvitsiani for discussions; L. McLean for assistance with recombinant proteins; R. Eifert, B. Burbach and R. Specht for technical support; S. Brenner-Morton for generating the guinea pig Fos antibody; K. Chatpar and Y. Sun for assistance with experiments; N. Zabello for help with mice; L. Stowers for the cat lipocalin (Fel-D4) plasmid; C. Denny for the gift of the Arc-CreER mouse; D. Hattori, J. Scribner, A. S. Lee and B. Noro for comments on the manuscript; and C. H. Eccard, P. Kisloff, A. Nemes and M. Gutierrez for laboratory support. This work was supported by the Howard Hughes Medical Institute (R.A.), the Biotechnology and Biological Sciences Research Council (J.L.H. and R.J.B.), and The Robert E. Leet and Clara Guthrie Patterson Trust Fellowship (E.D.).

**Author contributions** E.D., R.J.B., J.L.H., A.K. and R.A. discussed the design of experiments and the results, and wrote the manuscript. J.I.S. designed the custom behaviour and stimulation systems. E.D. performed all of the experiments and analysis. K.L. and N.B.-K. helped with the experiments and analysis. The recombinant MUPs were provided by R.J.B.

**Competing interests** The authors declare no competing interests.

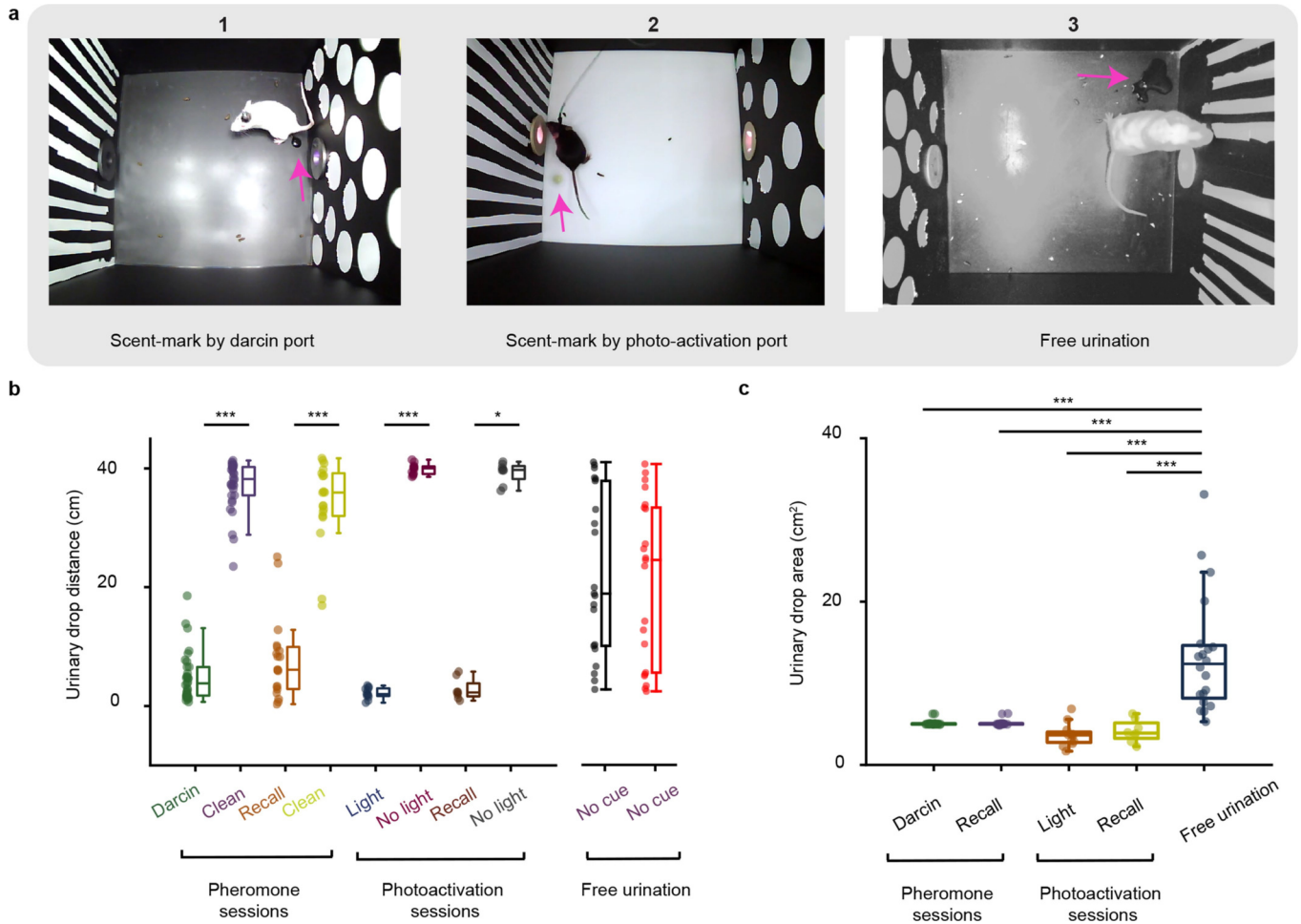
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1967-8>.

**Correspondence and requests for materials** should be addressed to A.K. or R.A.

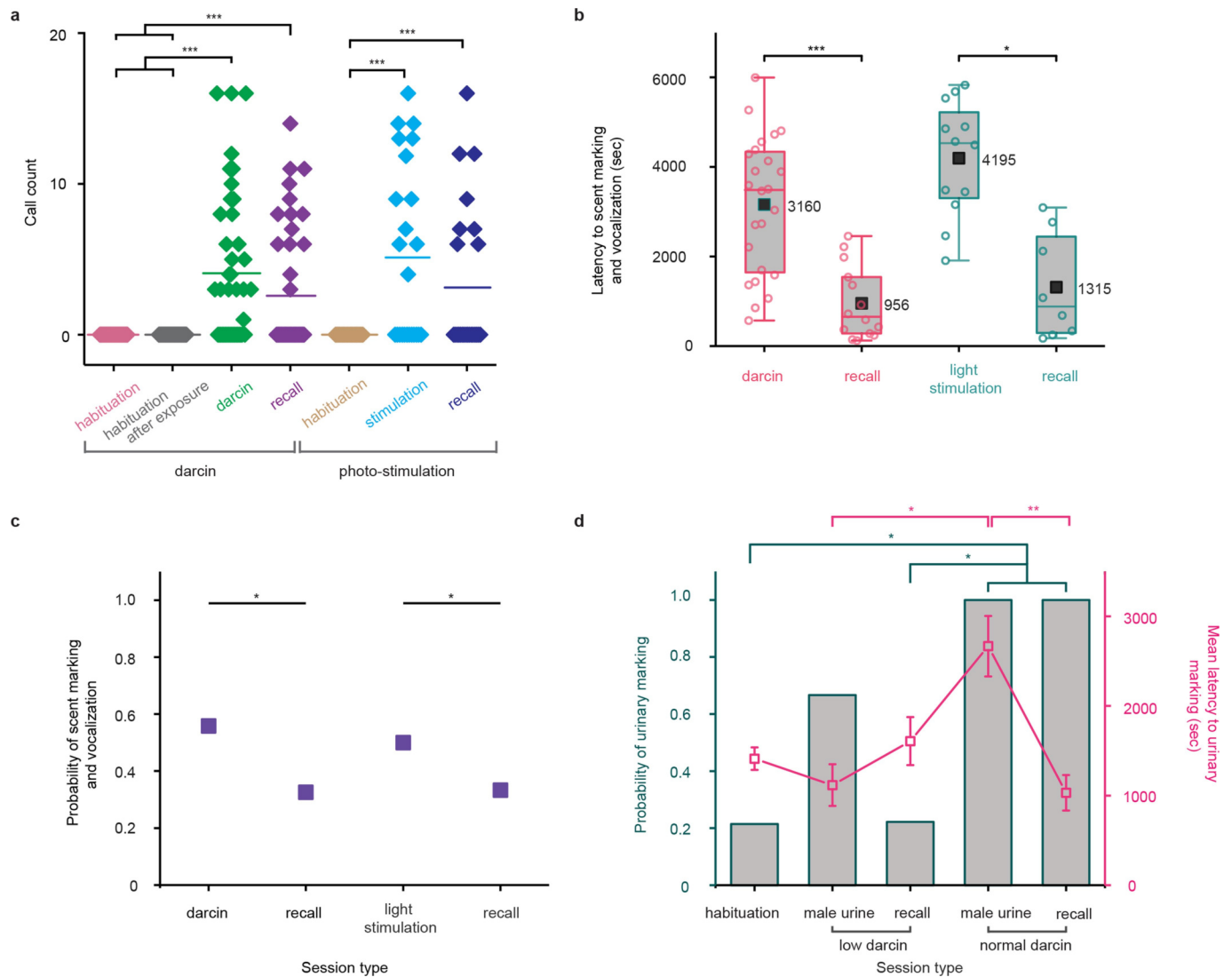
**Peer review information** Nature thanks Stephen Liberles and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



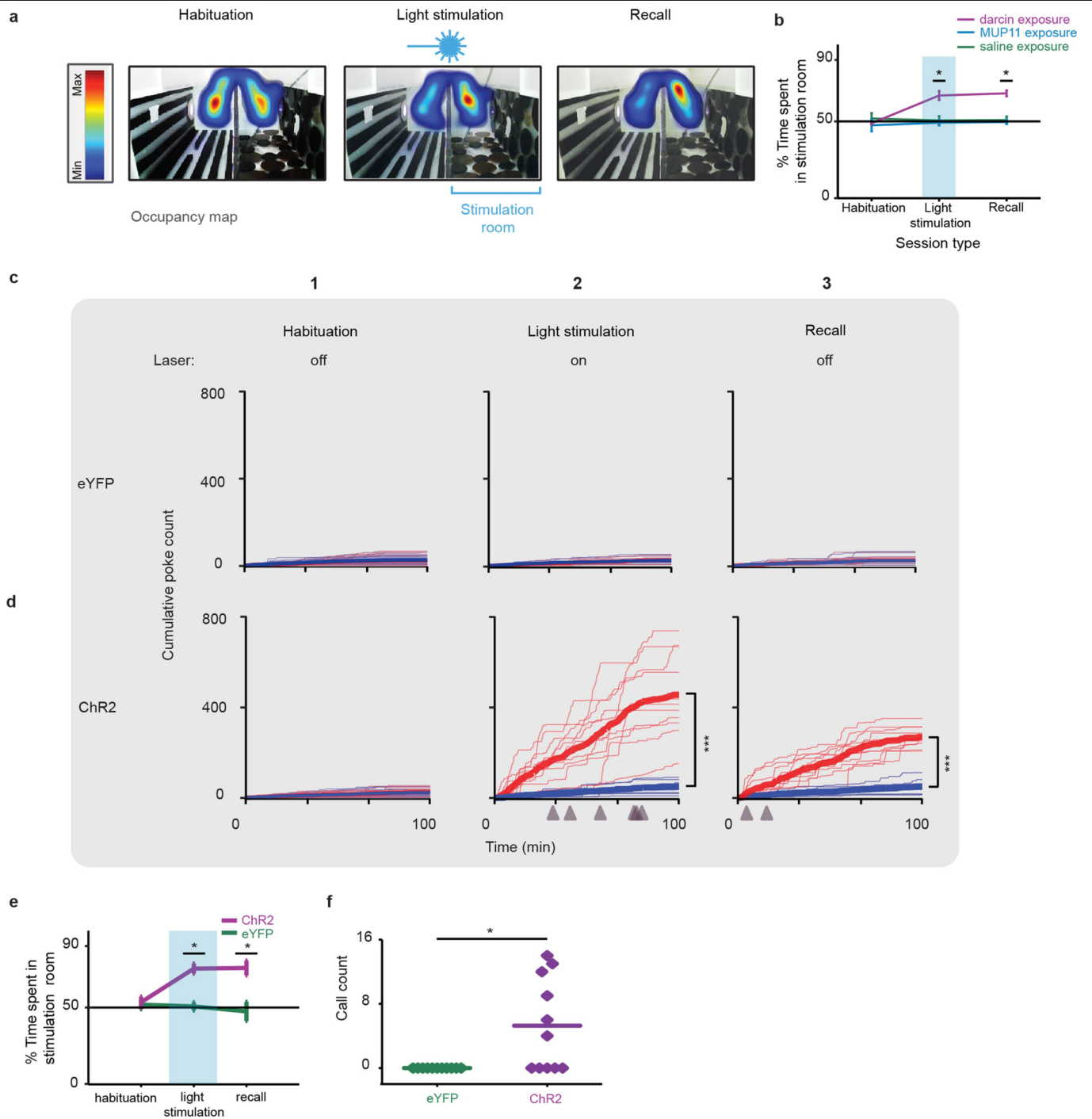
**Extended Data Fig. 1 | Darcin and photoactivation of posterior MeA neurons condition scent-marking place preference.** **a**, Representative frames from videos of the pheromone (1) and photoactivation (2) sessions, and free-range behaviours (3). **b**, Distance from urinary drop to each of the poke ports during various sessions. Individual frames were analysed using Adobe Photoshop CC to quantify the distance from the centre of a urinary drop to the base of each poke port. Units are scaled from pixels to centimetres. Distances were compared using the two-sided Wilcoxon signed-rank test (\*\*\* $P < 0.0005$ , \* $P = 0.01$ ;  $n = 24$  mice, (1);  $n = 12$  mice, (2);  $n = 20$  mice, (3)). **c**, Area of urinary drops under various conditions. Individual frames were analysed using Adobe

Photoshop CC to quantify the area of the urinary marks. Units are scaled from square pixels to square centimetres. Scent-mark area (mean  $\pm$  s.e.m., cm<sup>2</sup>): darcin,  $5 \pm 0.05$ ,  $n = 24$  mice; recall of darcin,  $5 \pm 0.09$ ,  $n = 14$  mice; photoactivation,  $4 \pm 0.4$ ,  $n = 12$  mice; recall of photoactivation,  $4 \pm 0.5$ ,  $n = 8$  mice; free urination,  $13 \pm 2$ ,  $n = 20$  mice. Areas were compared using the two-sided Mann-Whitney test (\*\*\* $P < 0.0005$ ), adjusted for multiple comparisons. The bounds in the box plots in **b**, **c** are defined by the 25th and 75th percentile of the distribution. The lines represent the median and the upper and lower whiskers represent the 75th percentile +  $1.5 \times$  IQR and 25th percentile -  $1.5 \times$  IQR, respectively.



**Extended Data Fig. 2 | Darcin and photoactivation of posterior MeA neurons reinforce recall of vocalization and scent-marking behaviours. a–c**, Data for individual mice for all unique sessions across the study were pooled. **a**, Mean (horizontal line;  $n = 43$  mice (darcin group),  $n = 24$  mice (photostimulation group)) and total calls made by individual mice (diamonds) detected during various sessions. Call counts were compared using the two-sided Wilcoxon signed-rank test within the respective groups ( $***P < 0.0005$ ), adjusted for multiple comparisons. **b**, Latency from the start of the session to urinary marking and vocalization behaviour (mean  $\pm$  s.e.m., seconds) during exposure to darcin ( $3,160 \pm 311$ ,  $n = 24$  mice), recall of darcin exposure ( $956 \pm 217$ ,  $n = 14$  mice), photostimulation ( $4,195 \pm 372$ ,  $n = 12$  mice) and subsequent recall ( $1,315 \pm 418$ ,  $n = 8$  mice) sessions. Latencies were compared within groups using the matched-pair two-sided  $t$ -test ( $*P = 0.005$ ,  $***P = 0.00009$ ). The bounds in the boxplots are defined by the 25th and 75th percentile of the distribution. The line represents the median and the upper and lower whiskers represent 75th percentile +  $1.5 \times$  IQR and 25th percentile –  $1.5 \times$  IQR, respectively. **c**, Probability of urinary scent-marking and vocalization behaviours. Mean probabilities are given for the darcin session ( $0.6$ ,  $n = 43$  mice), recall of darcin

session ( $0.3$ ,  $n = 43$  mice), photostimulation-evoked urinary marking and vocalization ( $0.5$ ,  $n = 24$  mice) and recall of photostimulation-evoked behaviours ( $0.3$ ,  $n = 24$ ). Probabilities were compared using the two-sided McNemar test ( $*P < 0.05$ ). **d**, Probability and mean latency to first urinary scent marking in the different sessions ( $n = 9$  mice). Data from 100-min habituation sessions (mean  $\pm$  s.e.m., latency for urination, seconds) and after exposure to male-soiled bedding in the home cage ( $1,411 \pm 126$ ), low-darcin urine from BALB/c mice ( $1,116 \pm 232$ ), recall session of BALB/c urine ( $1,607 \pm 268$ ), urine from C57BL/6J mice containing normal levels of darcin ( $2,666 \pm 337$ ) and recall of C57BL/6J urine ( $1,032 \pm 198$ ) are shown. Probabilities were compared using the two-sided McNemar test ( $*P = 0.02$ ), adjusted for multiple comparisons. Latencies were compared within groups using the matched-pair two-sided  $t$ -test and across groups using the unpaired two-sided  $t$ -test ( $**P = 0.0008$ ,  $*P = 0.02$ ), adjusted for multiple comparisons. Scent-marking behaviours in response to low-darcin urine during the subsequent recall sessions were compared (habituation to recall,  $P = 1$ , cue to recall session comparison,  $P = 0.1$ , two-sided McNemar test).

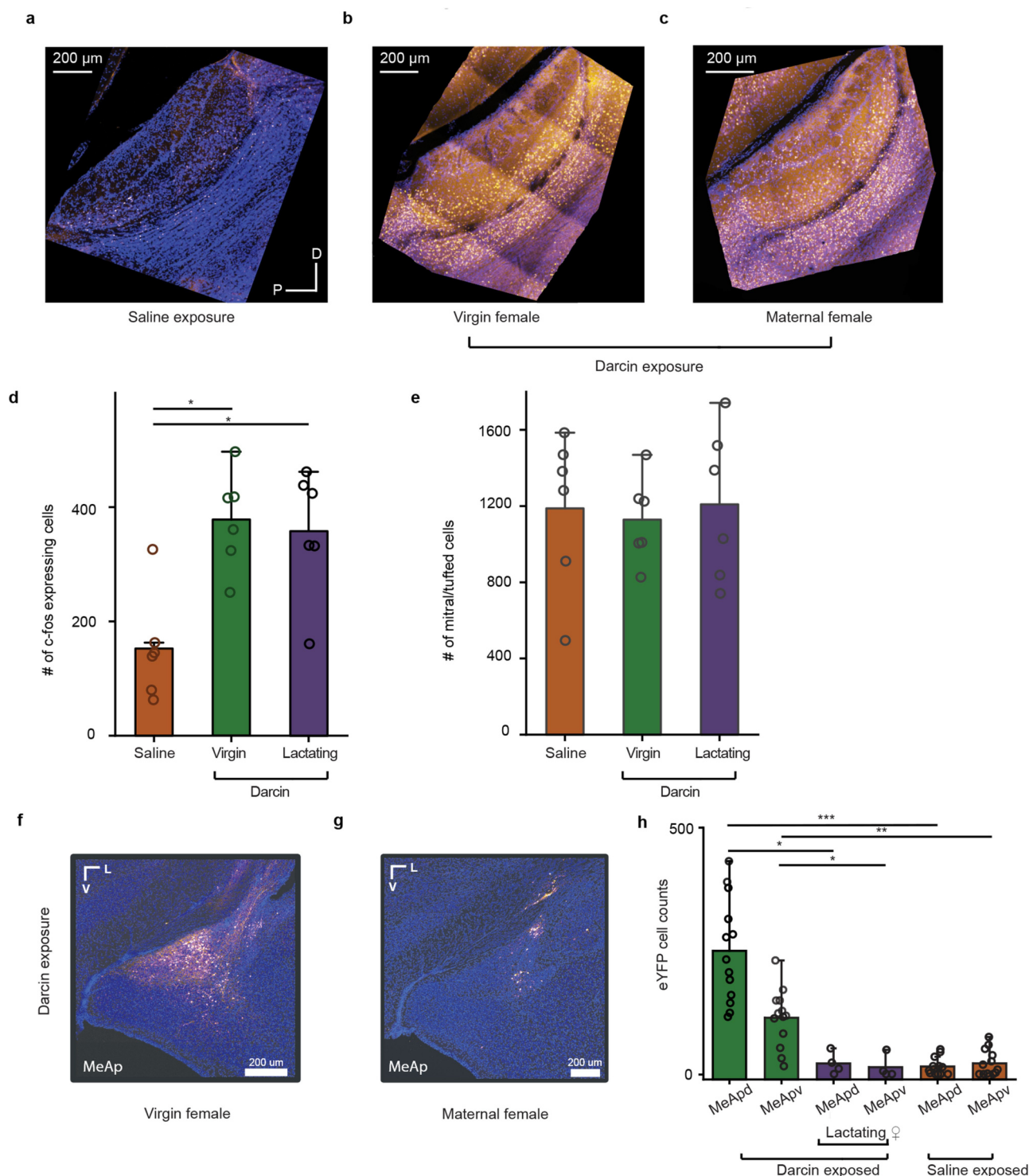


**Extended Data Fig. 3** | See next page for caption.



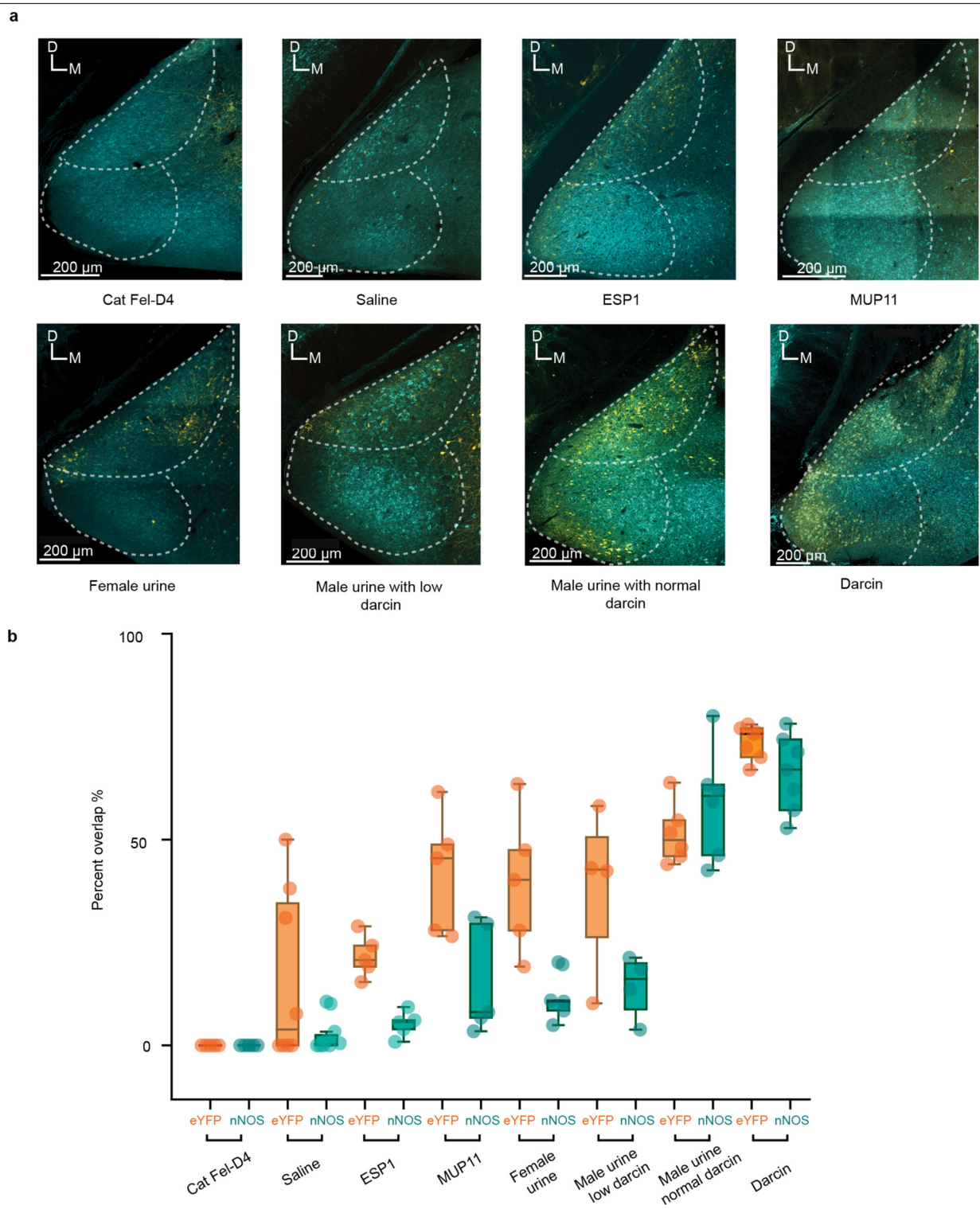
**Extended Data Fig. 3 | Activation of darcin-responsive neurons in the posterior MeA recapitulates darcin-induced behaviours.** **a**, Heat map showing occupancy of the chamber during a habituation, photostimulation and recall session. **b**, Occupancy plot showing the percentage of time spent in the photostimulation room. Arc-CreER mice were exposed to darcin (magenta), saline (green) or MUP11 (blue). The plot shows the mean  $\pm$  s.e.m. ( $n = 5$  mice per group, total  $n = 15$  mice) percentage of time spent in stimulation room during habituation, photostimulation and recall sessions. For occupancy time, pairwise comparisons were performed using the two-sided Mann–Whitney test ( $*P < 0.05$ ) and three-way comparisons were performed using Kruskal–Wallis tests (habituation  $P = 0.6$ , light stimulation  $P = 0.009$  and recall sessions  $P = 0.008$ ). **c–f**, Activation of nNOS neurons in the posterior MeA recapitulates darcin-induced behaviours. **c, d**, Cumulative poke counts during habituation (laser off; 1), light stimulation (laser on; 2), and recall (laser off; 3) sessions in mice expressing eYFP (**c**) or ChR2 (**d**) in nNOS neurons. Light stimulation was performed in one port (red) and not in the second port (blue). During habituation (1) and recall (3) sessions, no light stimulation was given, and red and blue reflect right and left ports, respectively. Mean (bold lines,  $n = 11$  mice

for each group) and individual (fine lines) cumulative poke counts are shown. The time-stamps for ultrasonic vocalization and scent-marking behaviours are indicated as arrowheads (**d** (2, 3)). Poke counts were compared using the two-sided Wilcoxon signed-rank test ( $***P = 0.0001$ ). Control group (eYFP) port entries (**c**) are contrasted to the ChR2 group (**d**) during light stimulation (red entries for ChR2 (**d** (2)) compared to eYFP (**c** (2));  $P = 0.0002$ ) and recall sessions (red entries for ChR2 (**d** (3)) compared to eYFP (**c** (3));  $P = 0.0002$ , two-sided Mann–Whitney test, adjusted for multiple comparisons). **e**, Occupancy plot showing the mean percentage of time spent in the photostimulation room by all mice during various sessions. nNOS-ires-Cre mice were injected with AAV encoding either eYFP (green) or ChR2–eYFP (purple); plots are colour-coded to their respective groups;  $n = 6$  mice per group,  $n = 12$  mice total. Occupancy times were compared using a two-sided Mann–Whitney test ( $*P < 0.05$ ). **f**, Mean (horizontal lines,  $n = 11$  per group,  $n = 22$  total) and total calls made by individual mice (diamonds) detected during the photostimulation (2) sessions in mice expressing eYFP (**c** (2)) or ChR2 (**d** (2)) in nNOS neurons. Call counts were compared using the two-sided Mann–Whitney test ( $*P = 0.007$ ).



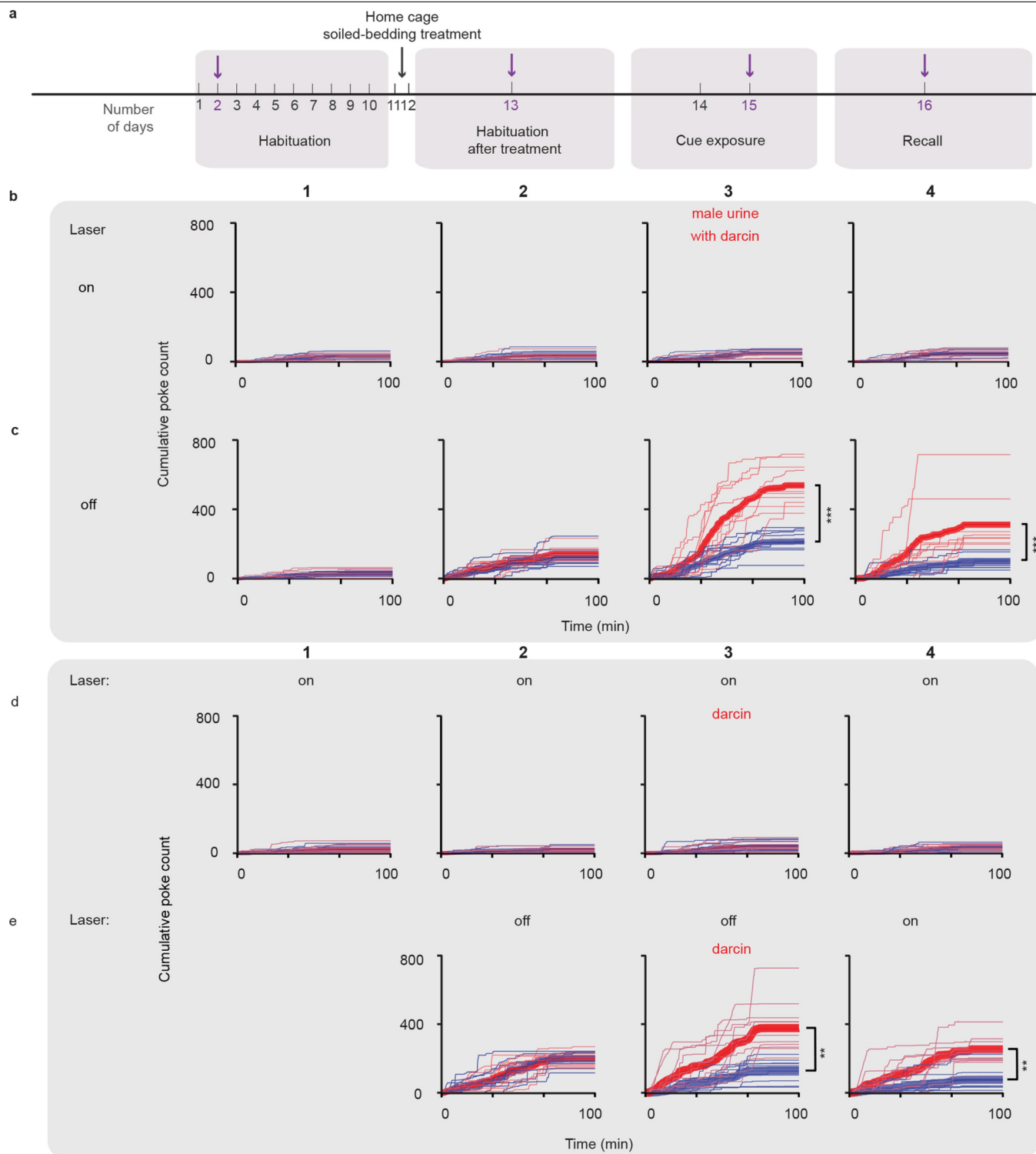
**Extended Data Fig. 4 | In lactating females, darcin activates mitral cells in the AOB but fails to activate MeA neurons.** **a–c**, Representative images showing Fos expression (orange) and NeuroTrace (blue) in sagittal sections of the AOB following exposure to saline (**a**) or darcin in virgin females (**b**) and lactating females (**c**). Experiment was independently repeated on 6 mice for each group. **d**, Bar plots quantifying Fos-expressing cells in the AOB. Fos counts (mean  $\pm$  s.e.m.): saline  $153 \pm 38$ , darcin in virgin females  $378 \pm 35$ , darcin in lactating females  $358 \pm 45$ ;  $n = 6$  mice per group. Cell counts were compared using the two-sided Mann–Whitney test ( $^*P = 0.02$ ), adjusted for multiple comparisons. **e**, Bar plots quantifying the mitral/tufted cells in the AOB. Number of cells (mean  $\pm$  s.e.m.): saline  $1,188 \pm 167$ , darcin in virgin females  $1,129 \pm 93$ , darcin in lactating females  $1,210 \pm 163$ ;  $n = 6$  mice per group. Cell

counts were compared using the two-sided Mann–Whitney test. **f, g**, Representative images showing eYFP expression in coronal sections of the posterior MeA of Arc-CreER mice after exposure to darcin in virgin females (**f**) and lactating females (**g**). Experiment was repeated on 13 mice and 4 mice in **f** and **g**, respectively. **h**, Bar plots quantifying eYFP-expressing cells in the MeApd and the MeApv. Cell counts were compared using the two-sided Mann–Whitney test, adjusted for multiple comparisons.  $^*P = 0.008$ ,  $^{**}P = 0.0006$ ,  $^{***}P < 0.0005$ . Mean  $\pm$  s.e.m. eYFP-expressing cell counts: saline,  $16 \pm 5$  in the MeApd and  $23 \pm 7$  in the MeApv,  $n = 13$  mice; darcin exposure in virgin females,  $251 \pm 29$  in the MeApd and  $115 \pm 16$  in the MeApv,  $n = 13$  mice; darcin exposure in lactating females,  $23 \pm 11$  in the MeApd and  $15 \pm 12$  in the MeApv,  $n = 4$  mice.



**Extended Data Fig. 5 | Identification of neurons in the posterior MeA that respond to vomeronasal stimuli and their overlap with the genetic marker nNOS. a,** Representative images showing the stimulus-responsive (eYFP, orange) and nNOS-expressing (cyan) neurons in the posterior MeA of Arc-CreER mice exposed to cat salivary lipocalin Fel-D4 ( $n = 5$  mice), saline ( $n = 8$  mice), ESP1 ( $n = 5$  mice), MUP11 ( $n = 5$  mice), female urine ( $n = 5$  mice), male urine with low levels of darcin ( $n = 4$  mice), male urine with normal levels of darcin ( $n = 9$  mice) and darcin ( $n = 7$  mice). **b,** Corresponding box plots

quantifying the percentage overlaps between the stimulus-responsive (eYFP) and nNOS<sup>+</sup> neurons in the posterior MeA of mice exposed to the various stimuli. Orange plots represent the percentage of YFP cells that overlap with nNOS; cyan plots represent the percentage of nNOS cells that overlap with YFP. The bounds in box plots are defined by the 25th and 75th percentile of the distribution. The lines represent the median and the upper and lower whiskers represent the 75th percentile +  $1.5 \times \text{IQR}$  and 25th percentile -  $1.5 \times \text{IQR}$ , respectively.

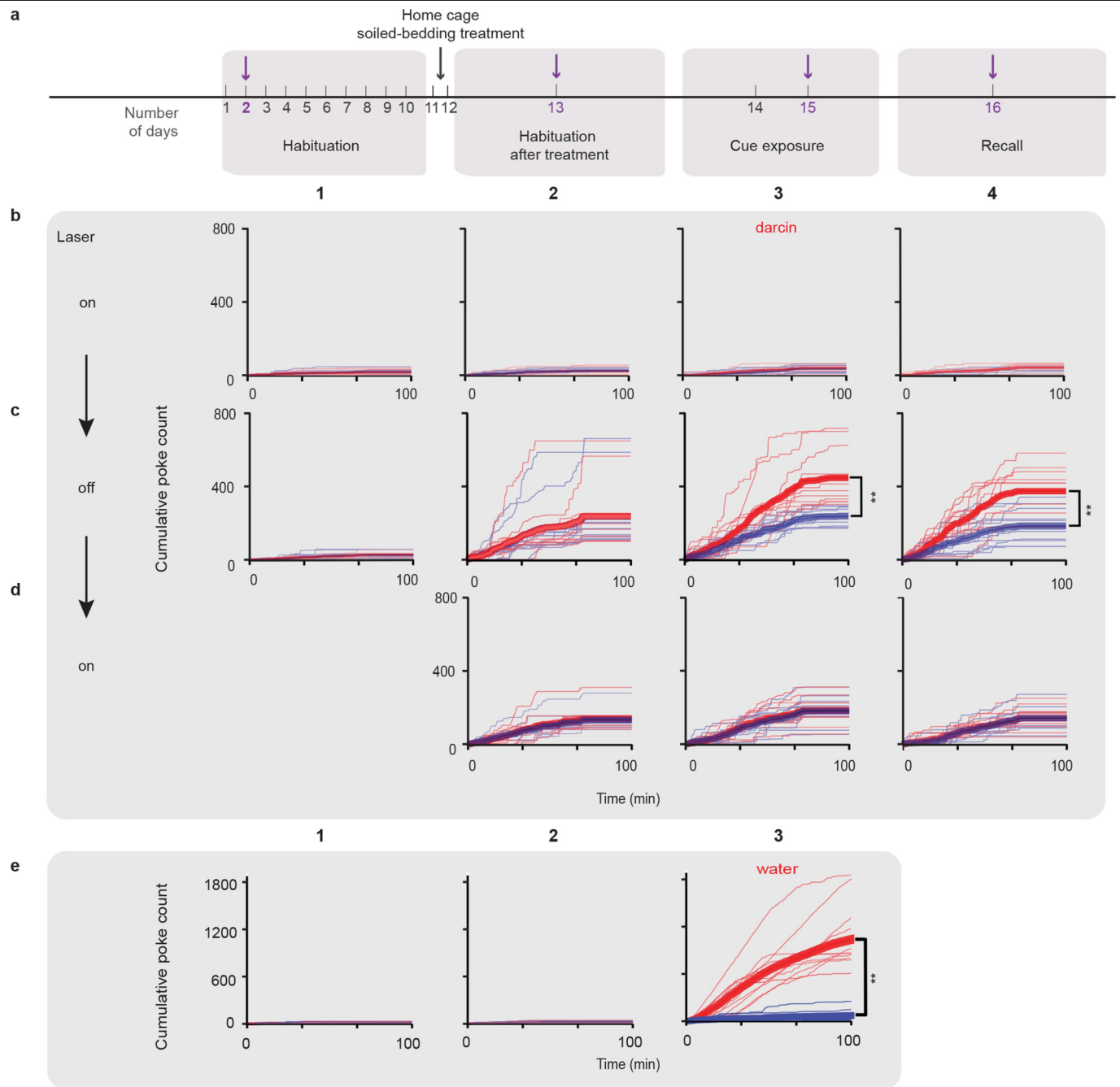


**Extended Data Fig. 6** | See next page for caption.

**Extended Data Fig. 6 | The additional effects of silencing nNOS neurons in the posterior MeA.** **a–c**, Functional convergence of both olfactory systems mediated by the posterior MeA is pivotal for male urine reinforcement. **a**, Timeline of the preference assay. Mice were habituated in the chamber for ten days (1), then exposed to male-soiled bedding for 60 h in their home cage (2), followed by one additional day of habituation before male urine (with normal levels of darcin ( $1 \mu\text{g } \mu\text{L}^{-1}$ )) was presented in one of the two ports (3). Urine was removed for the recall session one day later (4). Port preference was quantified from port entries. **b, c**, Cumulative poke counts during habituation (1), habituation after treatment (2), exposure to male urine (3) and recall (4) sessions for mice expressing eNpHR-eYFP ( $n = 10$ ) with (**b**) and without (**c**) optical silencing of nNOS neurons. Poke counts were recorded on the days indicated by purple arrows in **a**. Mice were exposed to male urine in one port (red) and a blank filter (blue) in the second port (3). During habituation (1, 2) and recall (4) sessions both ports contained a blank filter. Mean (bold lines) and individual (fine lines) cumulative poke counts are shown. Poke counts were compared using the two-sided Wilcoxon signed-rank test (\*\* $P = 0.0002$ ). The

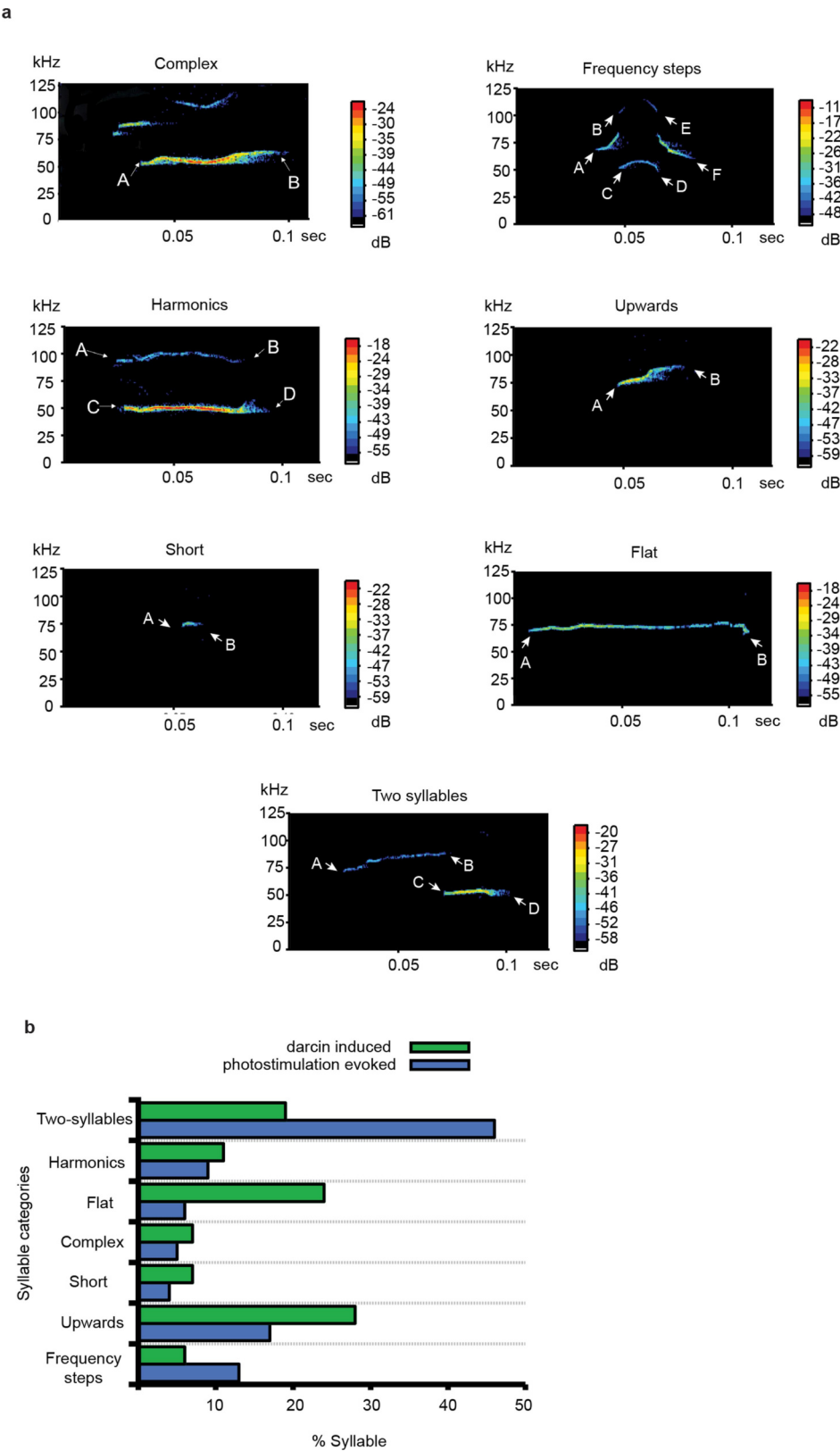
effect of silencing nNOS neurons is quantified with matched pair differences (male urine session comparisons, **b** (3) to **c** (3),  $P = 0.002$ ) and recall of male urine with darcin (recall session comparisons, **b** (4) to **c** (4),  $P = 0.002$ ) using the two-sided Wilcoxon signed-rank test, adjusted for multiple comparisons. **d, e**, Optical silencing of nNOS neurons does not affect recall of darcin memory. Cumulative poke counts during habituation (1), habituation after treatment (2), darcin (3) and recall (4) sessions in mice expressing eNpHR ( $n = 11$ ) with optical silencing during all sessions (**d** (1–4)) and with optical silencing during recall sessions only (**e** (4)). Poke counts were recorded on the days indicated by purple arrows in **a**. Mice were exposed to darcin in one port (red) and a blank filter (blue) in the second port (3). During habituation (1, 2) and recall (4) sessions both ports contained a blank filter. Mean (bold lines) and individual (fine lines) cumulative poke counts are shown. Poke counts were compared using the two-sided Wilcoxon signed-rank test (\*\* $P = 0.001$ ). The effect of silencing nNOS neurons during recall sessions was tested with matched pair differences (**c**, cue (**e** (3)) to recall (**e** (4)) comparisons, laser off (**e** (3)) and on (**e** (4)),  $P = 0.1$ , using the two-sided Wilcoxon signed-rank test, adjusted for multiple comparisons.





**Extended Data Fig. 7 | Mice subjected to optical silencing of nNOS neurons retained a motivation to poke.** To establish the primacy of the MeA in mediating darcin-evoked behaviours rather than altering general motivation, mice expressing eNpHR in nNOS neurons were also tested. **a**, Timeline of the two-port preference assay. **b–d**, Cumulative poke counts during habituation (1), habituation after exposure to male-soiled bedding in the home cage (2), darcin exposure (3) and recall (4) sessions with (b) and without (c) optical silencing of nNOS neurons, and with optical silencing again after 4 weeks (d) ( $n = 11$  mice). Poke counts were recorded on the days indicated by purple arrows in **a**. Mice were exposed to darcin in one port (red) and a blank filter (blue) in the second port. During habituation (1, 2) and recall (4) sessions both ports contained a blank filter. Mean (bold lines) and individual (fine lines) cumulative poke counts are shown. Poke counts were compared using the two-sided Wilcoxon signed-rank test (\*\* $P = 0.001$ ). The effect of silencing nNOS neurons after a learning experience is quantified during habituation sessions after exposure to soiled bedding in the home cage (port entries to the same port (red) with blank filters are compared during habituation after home-cage

treatment sessions in **b** (2) and **c** (2), laser on and off,  $P = 0.002$ , in **b** (2) and **d** (2), laser on,  $P = 0.001$ , and **c** (2) and **d** (2), laser off and on,  $P = 0.5$ ). The paired count differences (red–blue port) are compared across darcin sessions (**b** (3) to **d** (3), laser on,  $P = 0.5$ , and **c** (3) to **d** (3), laser off and on,  $P = 0.0001$ ) and recall of darcin (recall session comparison **b** (4) to **d** (4),  $P = 0.9$ , and **c** (4) to **d** (4),  $P = 0.0001$ ) using the two-sided Wilcoxon signed-rank test, adjusted for multiple comparisons. **e**, Optical silencing of nNOS neurons in the MeA does not affect non-social reinforcement behaviour. Cumulative poke counts during habituation (1), habituation after treatment (2), and water (3) sessions in mice expressing eNpHR ( $n = 12$ ) in nNOS neurons in the MeA with silencing. Poke counts were recorded on the days indicated by purple arrows in **a**. Water-restricted mice were rewarded with a drop of water ( $5 \mu\text{l}$ ) in one port (red) and a blank filter in the second port (blue). During habituation (1, 2) sessions both ports contained a blank filter. Mean (bold lines) and individual (fine lines) cumulative poke counts are shown. Poke counts were compared using the two-sided Wilcoxon signed-rank test (\*\* $P = 0.0005$ ).



**Extended Data Fig. 8 | Ultrasonic vocalizations that are emitted by mice exposed to darcin or stimulated optogenetically consist of seven unique syllable categories. a.** Representative spectrograms of ultrasonic vocalizations classified into seven categories of call. The heat maps show the intensities of the vocalizations. Descriptive statistics (mean  $\pm$  s.d., sample

sizes) for frequencies are given in Extended Data Table 2 for the locations indicated by the corresponding letters on the spectrograms. **b.** Percentages of different call categories emitted by mice exposed to darcin ( $n = 24$ , in green) and optogenetically stimulated ( $n = 12$ , in blue).

**Extended Data Table 1 | Cell counts for exposure to different cue types, nNOS expression and the overlaps in the posterior MeA**

Cue Type	YFP +	nNOS +	YFP + nNOS +	YFP +	nNOS +	YFP + nNOS +	YFP +	nNOS +	YFP + nNOS +	% nNOS / YFP
	in MeApd	in MeApd	in MeApd	in MeApv	in MeApv	in MeApv	in MeA pd+pv	in MeA pd+pv	in MeA pd+pv	% Overlap
Darcin (n=7)	200 ± 18	220 ± 22	146 ± 15	155 ± 21	179 ± 28	120 ± 21	311 ± 46	348 ± 54	232 ± 37	66 ± 3
Male urine normal darcin (n=9)	262 ± 45	155 ± 23	114 ± 16	168 ± 43	114 ± 25	63 ± 13	344 ± 33	275 ± 45	159 ± 21	59 ± 5
Male urine low darcin (n=4)	<b>102 ± 6*</b>	152 ± 17	<b>36 ± 12*</b>	64 ± 26	301 ± 123	21 ± 8	160 ± 14	250 ± 22	<b>41 ± 17*</b>	<b>22 ± 8*</b>
Female Urine (n=5)	<b>70 ± 12*</b>	160 ± 22	<b>27 ± 7*</b>	42 ± 4	172 ± 61	17 ± 5	<b>104 ± 15*</b>	297 ± 47	<b>45 ± 11*</b>	<b>11 ± 3*</b>
MUP11 (n=5)	<b>49 ± 10*</b>	206 ± 53	<b>17 ± 5*</b>	53 ± 12	183 ± 49	26 ± 8	<b>102 ± 17*</b>	390 ± 93	<b>43 ± 8*</b>	<b>16 ± 6*</b>
ESP1 (n=5)	<b>85 ± 18*</b>	323 ± 48	<b>14 ± 3*</b>	111 ± 30	576 ± 75	31 ± 8	197 ± 43	899 ± 93	<b>45 ± 11*</b>	<b>5 ± 1*</b>
Saline (n=8)	<b>22 ± 9*</b>	264 ± 41	<b>6 ± 3*</b>	<b>14 ± 8*</b>	290 ± 73	5 ± 4	<b>36 ± 15*</b>	554 ± 92	<b>11 ± 6*</b>	<b>2 ± 1*</b>
Fel-D4 (n=5)	<b>9 ± 9*</b>	204 ± 28	<b>0 ± 0*</b>	8 ± 8	189 ± 51	<b>0 ± 0*</b>	<b>13 ± 13*</b>	314 ± 91	<b>0 ± 0*</b>	<b>0 ± 0*</b>

Counts (mean ± s.e.m. cell counts) quantifying region-specific and overlapping expression of cue-responsive (eYFP<sup>+</sup> neurons) and nNOS-expressing neurons in the MeApd and the MeApv for female mice exposed to darcin, male urine with normal levels of darcin, male urine with low levels of darcin, female urine, MUP11, ESP1, cat Fel-D4 or saline. The percentage of overlap (mean ± s.e.m.) is quantified between total eYFP and nNOS-expressing neurons in the posterior MeA. Comparisons are made pairwise between darcin and all other cue types for eYFP<sup>+</sup> counts and the overlaps using the two-sided Mann–Whitney test (\**P* < 0.05). Comparisons are made pairwise between darcin and all other cue types for the percentage of nNOS-expressing neurons overlapping with eYFP using the two-sided Mann–Whitney test (\**P* < 0.05).

Extended Data Table 2 | Syllable categories for darcin and light-evoked ultrasonic vocalizations

	Darcin-Evoked	Light-Evoked	P(Tst)
	(mean ± sd)	(mean ± sd)	
Complex	n=5	n=9	
A	62kHz ± 23kHz	67kHz ± 2kHz	
B	68kHz ± 22kHz	72kHz ± 4kHz	
Duration	42ms ± 18ms	111ms ± 54ms	*
Harmonics	n=8	n=15	
A	83kHz ± 15kHz	66kHz ± 4kHz	*
B	91kHz ± 22kHz	78kHz ± 6kHz	
C	45kHz ± 8kHz	49kHz ± 3kHz	
D	46kHz ± 11kHz	49kHz ± 2kHz	
Duration	77ms ± 26ms	81ms ± 36ms	
Short	n=3	n=9	
A	75kHz ± 28kHz	74kHz ± 11kHz	
B	78kHz ± 26kHz	73kHz ± 11kHz	
Duration	8ms ± 4ms	8ms ± 3ms	
Two-syllable	n=45	n=27	
A	75kHz ± 8kHz	81kHz ± 8kHz	*
B	86kHz ± 12kHz	100kHz ± 6kHz	***
C	57kHz ± 10kHz	67kHz ± 6kHz	***
D	63kHz ± 9kHz	72kHz ± 8kHz	***
Duration	56ms ± 25ms	74ms ± 23ms	**
Upwards	n=17	n=39	
A	64kHz ± 12kHz	72kHz ± 9kHz	*
B	79kHz ± 8kHz	81kHz ± 12kHz	
Duration	19ms ± 9ms	36ms ± 17ms	***
Frequency steps	n=13	n=8	
A	74kHz ± 10kHz	77kHz ± 10kHz	
B	85kHz ± 9kHz	94kHz ± 14kHz	
C	52kHz ± 11kHz	62kHz ± 10kHz	*
D	53kHz ± 8kHz	64kHz ± 13kHz	*
E	86kHz ± 7kHz	100kHz ± 15kHz	*
F	76kHz ± 10kHz	91kHz ± 17kHz	*
Duration	71ms ± 26ms	97ms ± 36ms	
Flat	n=7	n=34	
A	62kHz ± 12kHz	70kHz ± 6kHz	
B	60kHz ± 13kHz	71kH ± 7kHz	
Duration	47ms ± 23ms	74ms ± 47ms	*

Different call categories emitted by mice exposed to darcin (n = 24) and optogenetically stimulated (n = 12). Frequencies and durations are compared with the unpaired two-sided t-test (\*P < 0.05, \*\*P < 0.005, \*\*\*P < 0.0006).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a                                 | Confirmed  |
|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

The behavioral nose poke data were acquired through a MATLAB 2009b-2013a (MathWorks Inc, Natick, MA) interface and an Arduino-powered device (Bpod v0.5, Sanworks, LLC, Stony Brook, NY). Avisoft Recorder USGH software was used to record vocalizations and integrate time codes from the Horita PTG2. The position of the mice was tracked using video tracking software XT10 (Noldus, Ethovision XT 10). A video recorder (AJA Ki Pro Recorder, AJA Video Systems, Grass Valley, CA, United States), which was connected to the camera and the Horita PTG2, was used to record video for the entire duration of the session. The time code generated by the Horita PTG2 was visible as an OLED (Organic Light Emitting Diode) display within the video window of the Marshall Electronics camera recording through the AJA recorder and was also recorded by screen-capturing software (Adobe Captivate CC 2017, Adobe Systems Inc.). The behavioral hardware (valves for water delivery and port sensors) and the laser for optogenetic stimulation were controlled by custom MATLAB 2009b-2013a programs (MathWorks Inc, Natick, MA) and two Arduino-powered devices (Bpod v 0.5, Sanworks LLC and PulsePal, Sanworks LLC).

#### Data analysis

The behavioral nose poke data were analyzed by a custom program written in MATLAB 2009b-2013a (MathWorks Inc, Natick, MA). To analyze the nature of the ultrasonic vocalizations, Avisoft SAS Lab Pro Version 5.2.12 was used. A fast Fourier transformation (FFT) was applied to the recordings to generate spectrograms; the following parameters were used: FFT length of 256, Hamming window, time window overlap of 50%, frequency resolution of 977 Hz, and time resolution of 0.5 ms. The vocalizations were frequency modulated to a human-audible range using Avisoft SAS Lab Pro Version 5.2.12. All spectrograms were additionally parametrized using SAP 2011, MUPET and MATLAB 2009b-2013a software. To analyze the urinary scent marking behavior of the animals, Adobe Premiere Pro CC 2017 (Adobe Systems Inc.) was used to process the video recordings from each session. The position of the mice was tracked using video tracking software XT10 (Noldus, Ethovision XT 10), and the occupancy trajectories and time-spent in each chamber were computed for analysis. All statistical analyses were performed using R (R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria), OriginLab Pro 2017-2019 (OriginLab Corp., Northampton, MA), and MATLAB 2009b-2013a (MathWorks Inc, Natick, MA).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.



## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to determine the sample size prior to the study, but sample sizes were consistent with the research papers in the relevant field.
Data exclusions	In very rare cases, mice that are confirmed to have no viral expression after behavioral testing were excluded, and these exclusion criteria were predetermined.
Replication	All attempts to replicate the behavioral, opto-genetics, and immunohistochemical experiments were successful, indicating the robustness of the presented results. To guarantee reliable replication of our results, we preferred to work with high infectious titer (3e6 IU/ml) and DJ serotype AAV viruses (as purchased from Stanford virus core), especially when we delivered the viruses into medial amygdala.
Randomization	Animals were randomly assigned to the test. The social cue or light stimulation (for optogenetic activation) sides were evenly split in a random manner between two ports across the animals to control for any potential side bias.
Blinding	All behavioral experiments were scored by an individual blind to the social cue or light stimulation side and experimental design.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

### Antibodies used

All antibodies used in this study are commercially available and validated by the manufacturers, except the guinea pig c-fos antibody. The following primary antibodies were used: anti-GFP (rabbit polyclonal, 1:1000, Rockland, catalog #: 600-401-215, lot#: 28983, 33267) or chicken polyclonal 1:400, Aves Lab, catalog#: GFP-1020, lot#: GFP879484, GFP697986), anti-nNOS (rabbit polyclonal, 1:400, Invitrogen, catalog#: 61-7000, lot#: 1207899A, 987786A, 1578834A, 797629A), and anti-c-fos (goat and rabbit polyclonal, 1:500, Santa Cruz Biotechnology, catalog #: rabbit sc 52, lot#: B1115, goat sc 52-G, lot#: 10215, J1613, and K1109, and guinea pig polyclonal c-fos, with RRID#, AB\_2814707, generated by Susan Brenner-Morton at ZMBBI, Columbia University). Secondary antibodies are all used at 1:500 dilutions (alexa-594 goat anti-rabbit: Jackson ImmunoResearch, catalog#: 111-585-003, lot#: 135 626, 140268, alexa-594 goat anti-rat: BioLegend, Clone: Poly4054, catalog#: 405422, lot#: B262774, alexa-633 donkey anti-goat: Life Technologies, catalog#: A21082, lot#1711470, alexa-488 goat anti-rabbit: Jackson ImmunoResearch, catalog#: 111-545-006, lot#131752, alexa-488 goat anti-chicken: Jackson ImmunoResearch, catalog#:

103-545-155, lot#: 139170, alexa-488 donkey anti-rabbit: Invitrogen catalog#: A21206, lot#: 1981155, alexa-488 goat anti-guinea pig: Jackson ImmunoResearch, catalog#: 706-545-148, lot#: 127887, 143798, alexa-488 donkey anti-chicken: Jackson ImmunoResearch, catalog#: 703-545-155, lot#: 126602, alexa-594 donkey anti-rabbit: Invitrogen, catalog#: A21207, lot#: 1987293, and NeuroTrace alexa 640/660, Molecular Probes, catalog#: N21483, lot# 1656094).

## Validation

All antibodies are validated to react with corresponding mouse antigens.

Rabbit anti-nNOS: validated by Invitrogen and Thermofisher (<https://www.thermofisher.com/us/en/home/life-science/antibodies/invitrogen-antibody-validation.html?icid=ab-search-learning-ab-validation>), tested with the following applications, ELISA, Western Blotting, Immunohistochemistry (frozen), see the reference: Huang, PL. et. al. (1995) Nature 377: 239-242. CiteAb database reports 53 citations for this antibody.

Rabbit anti-GFP: validated by Rockland Immunochemicals ([https://rockland-inc.com/store/Antibodies-to-GFP-and-Antibodies-to-RFP-600-401-215-O4L\\_18562.aspx](https://rockland-inc.com/store/Antibodies-to-GFP-and-Antibodies-to-RFP-600-401-215-O4L_18562.aspx)), tested with ELISA, Western Blotting, immunohistochemistry, IF microscopy applications, and reported that no reactivity was observed against Human, Mouse or Rat serum proteins. CiteAb database reports 36 citations for this antibody.

Chicken anti-GFP: validated by Aves (<https://www.aveslabs.com/products/green-fluorescent-protein-gfp-antibody>), tested with Western blot, and immunohistochemistry. See the reference: Lu J. et. al. 2017(10):1377-1383 for both GFP antibodies. CiteAb database reports 702 citations for this antibody.

c-fos antibodies are validated by Santa Cruz Biotech. using immunohistochemical, Western Blotting and immunofluorescence. See the references: Choi GB. et. al. Cell 146(6):1004-15 and Root CM. et. al. (2014) Nature 515 (7526): 269-73. These c-fos antibodies are discontinued by the manufacturer, therefore we generated a guinea pig polyclonal c-fos antibody with the help of Susan Brenner-Morton at ZMBBI, Columbia University. We validated guinea pig c-fos by comparisons to commercially available c-fos antibodies from Santa Cruz. We performed immuno-histochemistry on free floating mouse brain sections with various c-fos antibodies and counted the number of cells stained by each antibody for comparisons in various mouse brain regions.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

### Laboratory animals

Experiments were conducted with 279 female mice between 6 and 30 weeks old. Mice were purchased at 4 weeks old. The mouse lines used were Arc-CreER (a gift from Christine Denny at Columbia University; also available from Jackson Laboratory, Jax stock #022357), ICR outbred (CD-1) wild-type mice (Harlan/Envigo), Ai14 (Rosa-CAG-LSL-tdTomato), nNOS-ires-CRE (Jax stock #017526), vGlut-ires-CRE (Jax stock #028863), Gad2-T2a-NLS-mCherry (Jax stock #023140). The nNOS-ires-CRE mice were crossed to ICR outbred mice (Harlan/Envigo) for 15 generations to exchange their genetic background to the ICR mice.

### Wild animals

This study did not involve any field captured animals.

### Field-collected samples

This study did not involve any field captured samples.

### Ethics oversight

All surgical and experimental procedures were done in accordance with the National Institute of Health's Guide for the Care and Use of Laboratory Animals and approved by the Cold Spring Harbor Laboratory and Columbia University Medical Center Institutional Animal Care and Use Committees.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Cell stress in cortical organoids impairs molecular subtype specification

<https://doi.org/10.1038/s41586-020-1962-0>

Received: 3 June 2019

Accepted: 26 November 2019

Published online: 29 January 2020

Aparna Bhaduri<sup>1,2,6</sup>, Madeline G. Andrews<sup>1,2,6</sup>, Walter Mancia Leon<sup>2</sup>, Diane Jung<sup>1,2</sup>, David Shin<sup>1,3</sup>, Denise Allen<sup>1,3</sup>, Dana Jung<sup>1,2</sup>, Galina Schmunk<sup>1,3</sup>, Maximilian Haeussler<sup>4</sup>, Jahan Salma<sup>5</sup>, Alex A. Pollen<sup>1,2</sup>, Tomasz J. Nowakowski<sup>1,3</sup> & Arnold R. Kriegstein<sup>1,2\*</sup>

Cortical organoids are self-organizing three-dimensional cultures that model features of the developing human cerebral cortex<sup>1,2</sup>. However, the fidelity of organoid models remains unclear<sup>3–5</sup>. Here we analyse the transcriptomes of individual primary human cortical cells from different developmental periods and cortical areas. We find that cortical development is characterized by progenitor maturation trajectories, the emergence of diverse cell subtypes and areal specification of newborn neurons. By contrast, organoids contain broad cell classes, but do not recapitulate distinct cellular subtype identities and appropriate progenitor maturation. Although the molecular signatures of cortical areas emerge in organoid neurons, they are not spatially segregated. Organoids also ectopically activate cellular stress pathways, which impairs cell-type specification. However, organoid stress and subtype defects are alleviated by transplantation into the mouse cortex. Together, these datasets and analytical tools provide a framework for evaluating and improving the accuracy of cortical organoids as models of human brain development.

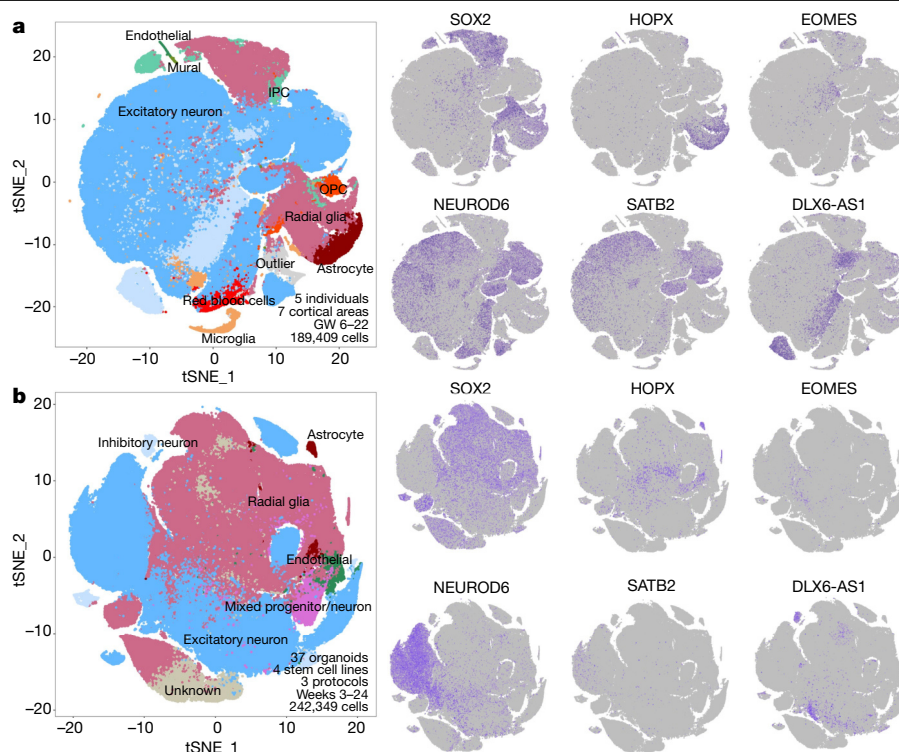
Organoid models harness the natural self-assembly properties of development to produce 3D cultures from stem cells that recapitulate aspects of an endogenous organ's structure and function. Organoids have applications in disease modelling, drug screening, and regenerative medicine. Single-cell RNA sequencing (scRNA-seq) provides a powerful method for comparing the fidelity of organoid cell types to their primary cell counterparts across tissues. In the liver and kidney, benchmarking comparisons with normally developing organs indicate that 3D culture better recapitulates primary cell types than adherent culture<sup>6,7</sup>. However, the lack of a comprehensive catalogue of cell types in the normal developing human brain and their molecular features has prevented careful evaluation of the strengths and weaknesses of cerebral organoids.

In vitro models of human cortical development are particularly valuable because early events during neurogenesis and synaptogenesis may underlie neuropsychiatric disorders, and experimental access to the developing human cortex is otherwise limited. Initial studies have indicated that broad classes of cells are preserved in cortical organoid models<sup>3,8</sup> but also hint at distinctions between organoids and primary cells<sup>4,9,10</sup>. In particular, the extent to which spatial and temporal gradients of gene expression and cell-type maturation are recapitulated in organoids is unclear (Extended Data Fig. 1). Although analysis of some of the first organoid models suggested the emergence of spatial gradients<sup>1,2,11</sup>, we know little about the fidelity and organization of areal cell types in organoids, in part because we lack molecular cell signatures of cortical areas in the developing brain.

## Comparison of human cortex and organoids

To evaluate the fidelity of cortical cell types in organoids, we performed high-throughput scRNA-seq of samples from the developing human cortex and cortical organoids, and compared the results with published organoid single-cell sequencing datasets. To characterize molecular features and gene-expression signatures during human cortical development, we performed scRNA-seq of dissociated cells from five cortical samples collected at 6–22 gestational weeks (GW), encompassing the period of neurogenesis. To assess cell-type differences across cortical areas, we studied primary samples from seven regions, including prefrontal (PFC), motor, parietal, somatosensory and primary visual (V1) cortices as well as hippocampus, resulting in transcriptomic data from 189,409 cells (Methods, Fig. 1a, Supplementary Table 1). These primary data were compared to data from 235,121 single cells from 37 organoids (Fig. 1b). We generated forebrain organoids by following three previously published protocols using different levels of directed differentiation to evaluate whether increased stringency in patterning signals results in more endogenous-like cellular subtypes<sup>1,4,8,12</sup> (Extended Data Fig. 2). To assess biological replicability, we used three induced pluripotent stem cell (PSC) lines and one embryonic stem cell line. Organoids were maintained under the same conditions, except for protocol-specific medium formulations (Extended Data Fig. 2), and samples were collected for immunohistochemistry and scRNA-seq after 3, 5, 8, 10, 15 and 24 weeks of differentiation to evaluate relevant cell types (Extended Data Figs. 3, 4). Last, we compared our reference

<sup>1</sup>Department of Neurology, University of California, San Francisco (UCSF), San Francisco, CA, USA. <sup>2</sup>The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco (UCSF), San Francisco, CA, USA. <sup>3</sup>Department of Anatomy, University of California, San Francisco (UCSF), San Francisco, CA, USA. <sup>4</sup>Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. <sup>5</sup>Center for Regenerative Medicine and Stem Cell Research, The Aga Khan University, Karachi, Pakistan. <sup>6</sup>These authors contributed equally: Aparna Bhaduri, Madeline G. Andrews. \*e-mail: [arnold.kriegstein@ucsf.edu](mailto:arnold.kriegstein@ucsf.edu)



**Fig. 1 | Cell types in cortical primary and organoid samples. a**, Single-cell sequencing of primary cortical cells identifies a number of cell types. These cell types are labelled in the *t*-distributed stochastic neighbour-embedding (*t*-SNE) plot on the left, and markers of cell-type identity depict progenitors (SOX2), oRGs (HOPX), IPCs (EOMES), newborn neurons (NEUROD6), maturing neurons (SATB2) and inhibitory interneurons (DLX6-AS1). Single-cell data can be explored at <https://organoidreportcard.cells.ucsc.edu>. **b**, Single-cell sequencing of cortical organoid cells generated from four different

pluripotent stem cell lines and three protocols with varied levels of directed differentiation generates similar cell types to primary cortex, but the population proportions differ. The proportion of cells for each marker in each sample type are: SOX2<sup>+</sup> (primary 15.4%, organoid 41.2%), HOPX<sup>+</sup> (primary 7.6%, organoid 4.2%), EOMES<sup>+</sup> (primary 4.1%, organoid 1.5%), NEUROD6<sup>+</sup> (primary 51.9%, organoid 20.3%), SATB2<sup>+</sup> (primary 32.5%, organoid 2.0%), and DLX6-AS1<sup>+</sup> (primary 17.1%, organoid 3.5%).

dataset to published organoid single-cell data generated from 276,054 cells across eight protocols, including time points from six months to a year<sup>3–5,8,9,13–15</sup>. This enabled us to extend our comparisons to later stages of differentiation (Extended Data Figs. 5, 6).

### Impaired cell-type fidelity in organoids

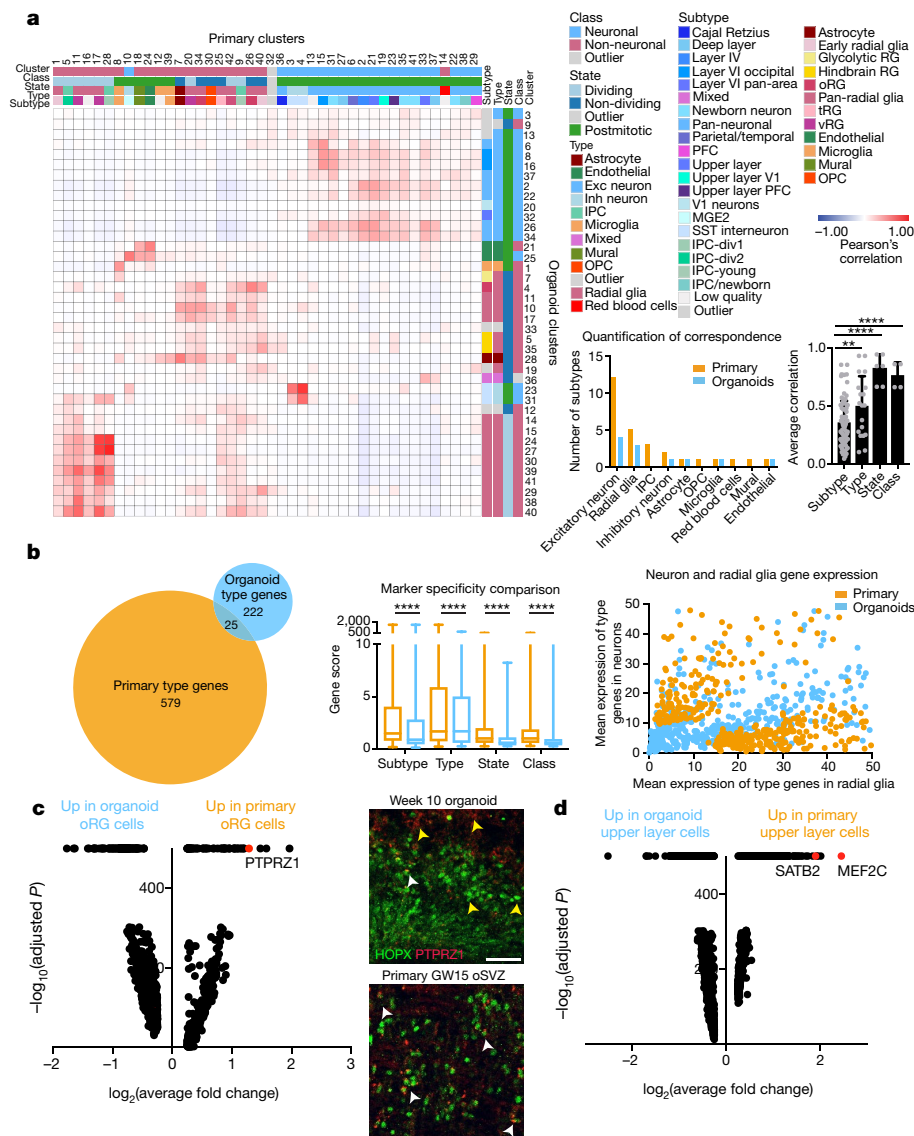
We identified broad cell types that corresponded to radial glia, intermediate progenitor cells (IPCs), maturing neurons and interneurons in both datasets (Fig. 1, Supplementary Tables 2, 3). In primary cortical samples, we also found clusters of microglia, oligodendrocyte precursors, mural cells and endothelial cells. Additionally, we identified previously described subtypes of radial glia, as well as a few instances of area- and layer-specific subtypes of excitatory neurons. In the primary samples, there was extensive intermixing within clusters of ages and cortical areas (Extended Data Fig. 4a). Within our organoids, cell lines, protocols and ages also intermixed, with variation primarily resulting from differences between cell types. Across lines and protocols, the forebrain marker FOXG1 was broadly expressed (Extended Data Figs. 5b, 7a), and the cell-type composition was similar across organoids of the same ages, even between lines and protocols, validating differentiation towards forebrain identity. Organoids had 45% fewer cells expressing HOPX, a marker of outer radial glia (oRG), than primary samples, and 63% fewer EOMES<sup>+</sup> IPCs, as previously noted<sup>3,16</sup>. We also found a 94% reduction in the number of SATB2<sup>+</sup> upper layer neurons in organoids compared to primary samples (Fig. 1b, Extended Data Fig. 3).

To quantitatively compare cell types in primary and organoid samples, we performed correlation analysis of marker genes (see Methods) based upon our clusters in each dataset. We categorized each

cluster in terms of its class (neuronal or non-neuronal), cell-cycle state (dividing or postmitotic), type (radial glia, excitatory neuron and so on) and subtype (for example, oRG, layer IV excitatory neuron), and quantified the correlation between organoid and primary cell categories. Neural class and proliferative state were largely preserved, as has been previously reported<sup>3,4,8,13</sup>. However, cell types ( $P=1.02 \times 10^{-20}$ ) and subtypes ( $P=5.34 \times 10^{-38}$ ) were significantly less well correlated to all organoid-derived cells, regardless of protocol (Fig. 2a). Our correlative analysis across all published organoid datasets suggested that a number of radial glia or neuronal clusters corresponded equally well to multiple primary cell subtypes and thus were designated as ‘pan-radial glia’ or ‘pan-neuronal’. Lack of subtype resolution resulted in a smaller number of high-quality subtypes in organoids compared to primary samples (see Methods; Fig. 2a). We validated our observation of limited subtype specificity between datasets using five additional batch correction methods and observed little overlap between organoid and primary clusters (Extended Data Fig. 8, Supplementary Table 5).

### Organoid radial glia lack specificity

The differentiation program that generates neurons from radial glia is highly conserved<sup>17–20</sup>, and we sought to identify genes that strongly discriminate progenitors from neurons. We were surprised to find that primary cell types are defined by more than twice as many genes as organoid cells, and that type-defining genes largely did not overlap between datasets (Fig. 2b, Extended Data Fig. 9, Supplementary Table 6). We used a gene score metric that quantifies the degree of enrichment and specificity for each marker gene in a dataset (Methods), which is initially low in primary cells but increases substantially over



**Fig. 2 | Molecular comparisons of cell subtypes between primary and organoid samples. a**, Each cluster was classified by marker genes for class, state, type and subtype (primary:  $n = 5$  individuals across independent experiments; organoids:  $n = 37$  organoids from 4 PSC lines across 4 independent experiments). Heat map shows correlation between pairwise combinations of marker genes (red intensity: Pearson's correlation from  $-1$  to  $1$ ). First histogram indicates cell subtypes in primary (orange) and organoid (blue) samples. Second histogram shows quantitative correlation from the best match for each category averaged across clusters (mean + s.d.; subtype versus type,  $**P = 0.0073$ ; subtype versus state,  $****P = 0.00008$ ; subtype versus class,  $****P = 0.003$ ; Welch's two-sided  $t$ -test). **b**, Marker specificity of primary and organoid cluster markers. Using VariancePartition, the genes defining metadata properties were evaluated for contribution to overall variance. Genes contributing  $>25\%$  variance by cell type were used in the Venn diagram. Box and whisker (mean + s.d.) depicts level of specificity at class ( $****P = 4.4 \times 10^{-14}$ ), state ( $****P = 4.7 \times 10^{-18}$ ), type ( $****P = 1.02 \times 10^{-20}$ ) and subtype

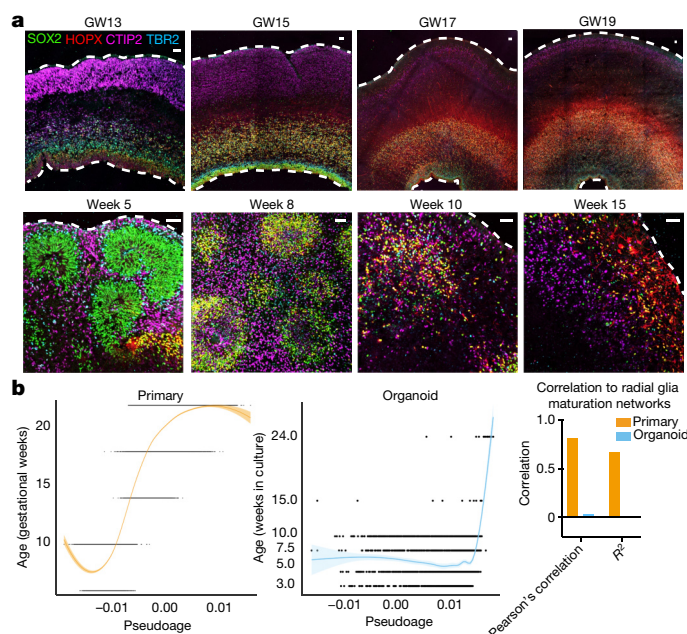
development (Extended Data Fig. 7e). In all cases, organoids exhibited a significantly lower gene score that did not resolve over time (Extended Data Fig. 7e), suggesting that markers of progenitors and differentiated cells might be co-expressed (Fig. 2b). We plotted the normalized counts for each gene that discriminated neurons from radial glia in primary samples, finding that neurons had low expression of radial glia markers, and radial glia did not express neuronal markers. However, we found substantial co-expression of these markers in organoid cells,

level ( $****P = 5.34 \times 10^{-38}$ ). Welch's two-sided  $t$ -test; primary,  $n = 5$ ; organoids,  $n = 37$ ). Dot plot depicts genes that discriminate between radial glia and neuron identity in primary samples. Each dot is a gene, shown as the average expression in radial glia (x-axis) and neuron (y-axis) for primary (orange) or organoid (blue) cells. **c**, Differential expression (two-tailed Wilcoxon rank sum test) between clusters annotated as oRG cells in primary and organoid datasets generated  $\log_2$ (fold change) (x-axis) and  $-\log_{10}$ (adjusted  $P$ ) (y-axis). Primary,  $n = 5$ ; organoids,  $n = 37$ ). A pseudocount of 500 was assigned to comparisons with an adjusted  $P = 0$ . Many measurements were significant, including expression of the oRG identity gene *PTPRZ1*. Week-8 organoids had minimal co-expression of *PTPRZ1* and *HOPX* (top), while at GW15 the outer subventricular zone (oSVZ) contains extensive co-localization (repeated independently 3 $\times$ ). White arrows, double-positive cells; yellow arrows, single-positive cells. Scale bar, 50  $\mu\text{m}$ . **d**, Differential expression (two-tailed Wilcoxon test) between cell clusters annotated as upper layer neurons.

resulting in a lower correspondence between organoid and primary cell types and subtypes.

We explored how well organoid radial glia recapitulated their primary cell subtype counterparts at the transcriptomic level by focusing the comparison on oRG cells. A number of genes were more highly expressed in organoid oRGs than in primary oRGs, and these genes were largely related to glycolysis or endoplasmic reticulum (ER) stress (Supplementary Table 7). One of the genes that was most highly





**Fig. 3 | Maturation of cortical lamina and radial glia. a.**

Immunohistochemistry of SOX2<sup>+</sup> progenitors, HOPX<sup>+</sup> oRG cells, CTIP2<sup>+</sup> deep layer neurons and TBR2<sup>+</sup> IPCs in primary and organoid samples showing laminar structure during neurogenesis. Primary samples express SOX2 and TBR2 in the ventricular zone and CTIP2 in the cortical plate at GW13. By GW15, HOPX<sup>+</sup> oRGs are born and reside in the oSVZ. The cortex expands markedly over the following weeks with more HOPX<sup>+</sup> oRG cells residing in the oSVZ, providing a scaffold on which neurons migrate. Organoids express similar markers to GW13 samples by week five of differentiation with multiple ventricular zone-like structures. oRG cells arise and increase between weeks eight and ten. The radial architecture expands and dissolves over this period. By week 15, a mix of cell types is present in the organoid. Organoids shown were differentiated using the 'least directed' differentiation protocol and staining was validated independently three times (primary:  $n = 4$  biologically independent samples; organoid:  $n = 3$  biologically independent samples). Scale bar, 50  $\mu$ M. **b.** Pseudotime was calculated by identifying networks from a 10,000-cell subset of primary radial glia that highly correlated to age (either positively or negatively). These networks were then collapsed into a single 'age network'. The module eigengene for this age network was then calculated on the remaining data and used for pseudotime. Pseudotime is indicated by the graph line and shading represents the geometric density standard error of the regression. The primary dataset (orange) has a high Pearson's correlation and  $R^2$  value, while the organoid dataset has no correlation to the pseudotime metric.

upregulated in primary oRG cells, but had very low expression in organoid oRG cells, was *PTPRZ1*, a known oRG marker<sup>21</sup> (Fig. 2c). To validate this finding, we stained primary and organoid samples for *PTPRZ1* and HOPX, a canonical oRG marker<sup>22,23</sup>, and found that primary tissue contained more HOPX and *PTPRZ1* co-expressing cells than did organoids (Fig. 2c, Extended Data Fig. 9e). We performed similar differential expression analysis between upper layer neuron clusters, and found that two genes required for neuronal maturation and projection pattern specification, *MEF2C* and *SATB2*<sup>24,25</sup>, were substantially upregulated only in primary cells (Fig. 2d). Even when cellular subtypes can be assigned to organoids, they lack molecular subtype identifiers.

### Cell maturation is impaired in organoids

The progression of developmental events, such as the birth of neuronal and glial cell types, occurs more rapidly in organoids than in primary tissue, and progenitor and neuronal zones do not expand as broadly as in vivo (Fig. 3a). A primitive radial glial scaffold is observed

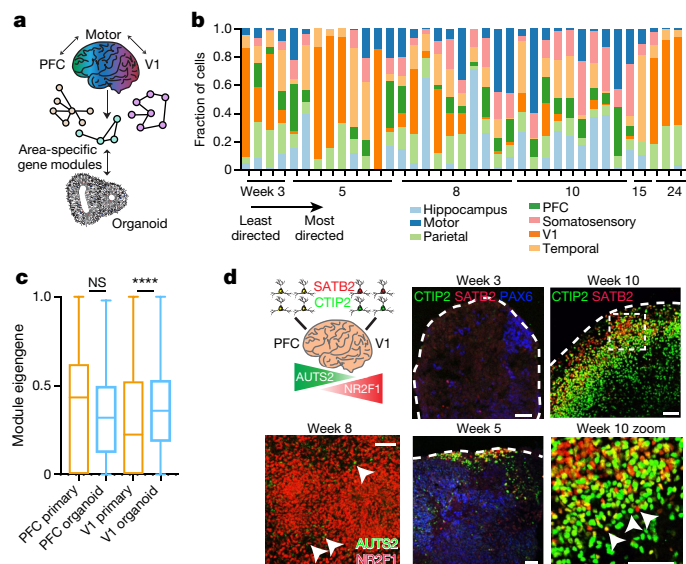
at 5 weeks of differentiation in organoids, whereas the oRG scaffold expands predominantly after 15 weeks of development in the primary human cortex. Over the course of 15 weeks of differentiation, the organoid progenitors differentiate, the 'scaffold' dissolves and intermixed populations of neurons and glia develop. Using our transcriptomic data, we sought to explore how cellular maturation was affected as a result of the faster temporal development we observed cytoarchitecturally in organoids. To explore the maturation of progenitor cells, we used weighted gene co-expression network analysis (WGCNA<sup>26</sup>) to generate gene modules of strongly correlated genes in primary radial glia. We consolidated networks that correlated with sample age into a pseudotime metric and then correlated pseudotime with actual age, observing a strong positive correlation in primary radial glia (Fig. 3b, Extended Data Fig. 10, Supplementary Table 8). With confidence in our networks, we applied them to the organoid radial glia. We saw limited correlation between organoid pseudotime and actual age, suggesting that the molecular maturation programs that exist in vivo are not activated in organoids. Notably, this heterogeneous maturation level existed within each organoid (Extended Data Fig. 10c), indicating that there is variability between individual cells and not just across organoids, lines or batches. The lack of a radial glia molecular maturation signature in organoids correlates with the absence of molecular diversity in this model. The effect of radial glia subtype and maturation on the role of these cells as neural precursors is unknown, but the dysregulation of these programs in organoids may affect their ability to completely recapitulate differentiation trajectories of cortical neurons in vivo.

### Definition of cortical areal signatures

Recent studies have uncovered molecular differences between excitatory neurons across cortical regions<sup>27–29</sup>, and these differences may emerge during neurogenesis<sup>19</sup>. Given that regional specification may represent a central feature of neuronal identity, we investigated how the molecular properties of areal identity emerge. We leveraged primary cell data collected from seven cortical regions. For genes that were uniquely enriched in each region, we calculated a weighted average expression (eigengene) across primary and organoid cells (Supplementary Table 9). In primary cells, some signatures, such as those from the PFC, temporal lobe, hippocampus and V1, were highly enriched in their respective areas (Extended Data Fig. 11). Notably, the parietal lobe tracked closely with the temporal lobe and the somatosensory and motor cortex co-expressed signatures, suggesting a lack of areal segregation between these regions at the time points sampled (Extended Data Fig. 11c). The earliest samples in our dataset preceded the development of anatomical distinctions between cortical regions, and thus could not be subdivided. The early 'telencephalon' samples were highly enriched for V1 signatures, but additional work is required to clarify whether excitatory neurons born early in development all begin by expressing V1 areal genes, or if this was a sampling artefact of our dissections. These data offer a new categorization of cortical area signatures and enable us to evaluate the areal identities of cortical organoid neurons.

### Areal signatures reflected in organoids

Our analysis indicates that many aspects of neuronal subtype are not preserved or are averaged into a pan-neuronal identity in organoids. However, our primary data suggest that areal identity is an early marker of neuronal differentiation. Using areal signatures from primary cells, we were able to evaluate the closest areal identity for each excitatory organoid neuron profiled by scRNA-seq. We were surprised to discover that most neurons corresponded to a defined areal signature (Fig. 4b) despite the lack of thalamic input, which is thought to refine areal identity<sup>30,31</sup>. Although each organoid contained neurons with multiple areal identities, the strength of areal correspondence of organoid excitatory



**Fig. 4 | Analysis of areal identity in organoid excitatory neurons.** **a**, Each of seven cortical areas was used to generate a unique area gene signature by comparing expression with the other six areas. The unique signatures were considered networks, and module eigengenes across area networks were calculated for each primary and organoid cell. The area with the highest normalized eigengene (normalized to the highest score within each area for equal comparison) was designated as the areal identity of that cell. **b**, Cortical composition for organoids across protocols. Areal identity was assigned for each cell within an organoid and the areal composition is shown for the 37 organoids in our dataset. Organoid samples are listed from earliest to latest stage collected (weeks 3–24). Within a time point, the organoid protocol used is ordered from least to most directed differentiation; each time point is comprised of multiple PSC lines. Every organoid has heterogeneous areal expression. **c**, The average module eigengene score for each primary (orange) and organoid (blue) cell designated (primary) or assigned (organoid) PFC or V1 identity (primary,  $n = 5$  independent samples across 5 experiments; organoid,  $n = 37$  organoids across 4 independent experiments). The average value for PFC was not significantly different between organoid and primary, and the V1 organoid cells had higher correlation to the V1 signature than primary cells, indicating that areal identity in the organoid strongly resembles normal development (box plots: centre line shows mean, box limits show range and whiskers show standard deviation; two-sided Welch's  $t$ -test,  $P = 0$ ). **d**, Validation of intermixing of areal identities in organoid samples differentiated using the least directed differentiation protocol. In the PFC, BCL11B and SATB2 co-localize in the same cell, whereas in V1 cells they are mutually exclusive. Both patterns are in close proximity in the organoid. AUTS2 is a rostrally expressed transcription factor whereas NR2F1 is a caudally expressed factor, but they are adjacent in the organoid. Scale bar, 50  $\mu$ m; representative image shown ( $n = 3$  replicates each).

neurons was robust, including to regions such as PFC and V1 (Fig. 4c, Extended Data Fig. 11d). Regardless of the PSC line or differentiation protocol used, cortical organoids comprised heterogeneous areal identities. To investigate whether cells that corresponded to different areas were spatially segregated within an organoid, we performed immunohistochemistry for two sets of area-specific genes. PFC excitatory neurons co-express the projection-specification transcription factors SATB2 and BCL11B (also known as CTIP2), and through a narrow topographical transition these markers segregate entirely in V1<sup>19</sup>. We explored the expression of these factors in our organoids, and observed both co-expression and segregation of SATB2 and CTIP2 in adjacent cells (Fig. 4d). *AUTS2* and *NR2F1* are well-described genes with rostral–caudal gradient expression patterns, and we similarly observed cells expressing either of these factors in proximal space. Together, these data suggest a model in which differentiation of cortical excitatory

neurons is strongly defined by areal identity, and organoids recapitulate this process, but without spatial organization.

### Cellular stress increases in organoids

Modules related to the activation of glycolysis and ER stress are enriched in organoid cells<sup>4</sup>, and additional analysis using four orthogonal co-clustering methods showed that stress pathways were upregulated in organoids across all protocols (Extended Data Figs. 12, 13). We confirmed that several genes that were upregulated in organoid datasets<sup>4</sup>, including the glycolysis gene *PGK1*<sup>32</sup> and the ER stress genes *ARCN1*<sup>33</sup> and *GORASP2*<sup>34,35</sup>, were enriched at the protein level in organoids (Extended Data Fig. 12). Immunostaining verified that, regardless of the stage of organoid differentiation or PSC line used, there was increased expression of PGK1, ARCN1 and GORASP2 in distinct organoid domains, not restricted to the organoid core.

To probe the origin of this cellular dysregulation, we first evaluated the expression of stress genes during normal human cortical development. Fixed cryosectioned tissue samples showed little expression of these genes throughout peak neurogenesis (Extended Data Fig. 12c), though some ER stress was observed at earlier cortical stages (Extended Data Fig. 13b, c). As ER stress and glycolysis genes are not canonically activated during cortical development, we hypothesized that the *in vitro* conditions of the organoid model resulted in increased cellular stress. We first evaluated the activation of stress pathways in PSCs and were surprised to find expression of both *ARCN1* and *GORASP2*, suggesting that ER stress occurs in stem cells before organoid formation (Extended Data Fig. 12b). To determine the rate at which cellular stress arises *in vitro*, we cultured organotypic cortical slices for one week and observed negligible change in stress activation compared with acutely fixed samples. However, we did observe upregulation of *ARCN1* and *GORASP2* in primary dissociated cells after one week in culture (Extended Data Figs. 12, 13), suggesting that cellular stress may be a broader feature of *in vitro* culture.

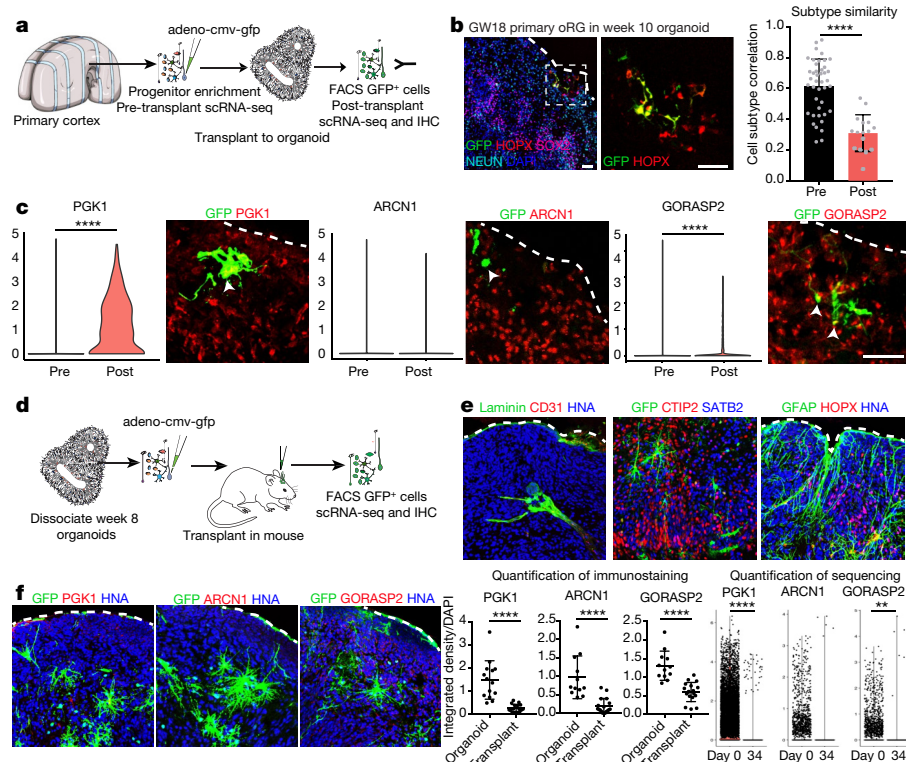
### Organoid environment activates stress

To test whether aggregate cell-culture conditions induced cellular stress, we transplanted GFP-labelled primary progenitors from GW14–20 into organoids. After 2.5 weeks we observed GFP-labelled SOX2<sup>+</sup> HOPX<sup>+</sup> primary radial glia within the organoids (Fig. 5b, Extended Data Fig. 14). We isolated GFP<sup>+</sup> primary cells and performed scRNA-seq to compare pre- and post-transplanted cells to our primary reference dataset (Supplementary Table 10). We found a marked increase in the expression of the glycolysis gene *PGK1* and the ER stress gene *GORASP2* in primary cells transplanted into, or generated within, organoids (Fig. 5c, Supplementary Table 11).

### Increased stress impairs cell subtype

Mouse knockout studies have suggested that the activation of ER stress pathways can inhibit cell-type specification<sup>36,37</sup>, so we investigated whether metabolic stress affected specification in transplanted primary cells. We noted similar subtypes when we compared pre-transplantation cell clusters and our primary reference data. By contrast, post-transplanted primary cells had significantly lower subtype correlation, similar to organoid cells, and they lacked markers of specific progenitor or neuronal subtypes (Fig. 5b, Extended Data Fig. 14e). To test whether the differences were driven by the induction of stress pathways, we also generated 3D aggregates of dissociated primary cortical cells from GW14/15 and found a similar upregulation of metabolic stress genes (Extended Data Fig. 13e). However, we observed an intermediate phenotype in which subtype correlation was significantly ( $P = 0.0037$ ) higher in primary aggregates than in post-transplanted primary cells, but still significantly lower than in pre-transplant cells (Extended Data





**Fig. 5 | Influence of culture on metabolic stress and cell type.** **a**, Primary samples were progenitor-enriched, GFP-labelled and transplanted into organoids. After 2.5 weeks, GFP<sup>+</sup> cells were isolated by fluorescence-activated cell sorting (FACS) and processed for scRNA-seq. IHC, immunohistochemistry. **b**, Primary GFP<sup>+</sup> cells integrate into organoids differentiated using the least directed protocol. Scale bar, 50  $\mu$ m. Pre-transplantation cells have similar profiles and cellular subtypes as primary data. After transplantation, there is a decrease in subtype correlation ( $n = 7$  biologically independent samples across 2 independent experiments). Mean  $\pm$  s.d. subtype correlation indicated on graph ( $P = 2.8 \times 10^{-9}$ , two-sided Welch's  $t$ -test). **c**, After transplantation, primary cells show increased expression of the stress genes *PGK1* ( $****P = 1.76 \times 10^{-87}$ , two-sided Student's  $t$ -test) and *GORASP2* (arrow;  $****P = 9.60 \times 10^{-63}$ , two-sided Student's  $t$ -test) as indicated by width of coloured domain in each respective violin plot ( $n = 7$  samples across 2 experiments). Scale bar, 50  $\mu$ m. **d**, Organoids were dissociated, GFP-labelled and injected into the cortex of P4 mice. After

2–5 weeks, mouse brains were collected for scRNA-seq and immunostaining. **e**, Human cells are visualized by GFP and human nuclear antigen expression ( $n = 13$  mice transplanted with 14 organoids derived from 2 induced PSC lines across 2 independent experiments). Organoid cells express markers of progenitors (HOPX), neurons (CTIP2 and SATB2) and astrocytes (GFAP). Mouse vascular cells (laminin and CD31) innervate the transplant. **f**, Post-transplantation, organoid cells show reduced expression of stress genes. Scatter plots show decreased staining intensity in transplanted organoids (error bars, s.d.) of *PGK1* ( $****P = 9.64 \times 10^{-7}$ ), *ARCNI* ( $****P = 1.28 \times 10^{-3}$ ) and *GORASP2* ( $****P = 1.88 \times 10^{-6}$ ;  $n = 19$  sections from 6 transplanted mice across 2 experiments, each marker stained independently; two-sided Student's  $t$ -test). Violin plots show a decrease in *PGK1* ( $****P = 2.16 \times 10^{-17}$ ) and *GORASP2* ( $**P = 0.0019$ ) expression in organoid cells post-transplant from single-cell analysis ( $n = 1,980$  cells from 7 transplanted mice across 2 experiments, two-sided Welch's  $t$ -test).

Fig. 14e). Some of these discrepancies might be attributable to the presence of other cell types, including microglia, endothelial cells and pericytes in the primary aggregate, that may promote normal maturation and differentiation.

## Transplantation rescues cell stress

To determine whether an in vivo environment could rescue the cellular stress derived from organoid culture conditions, week-8 organoids were dissociated, virally labelled with GFP and transplanted into the cortices of postnatal day four (P4) mice (Fig. 5d). Two or five weeks after transplantation, organoid-derived cells could be visualized incorporated into the mouse cortex (Fig. 5e, Extended Data Fig. 14f). After five weeks, organoid-derived cells had intricate morphologies and showed reduced expression of cellular stress markers. The glycolysis gene *PGK1* and the ER stress gene *ARCNI* were not expressed, and the ER stress gene *GORASP2* showed reduced expression compared to normal organoid conditions (Fig. 5f, Extended Data Fig. 14g). As organoid cells showed reduced stress after transplantation, we evaluated whether organoid-derived cells were capable of higher subtype specificity when removed from the in vitro environment. We isolated GFP<sup>+</sup> organoid cells two or five weeks after transplantation for scRNA-seq and compared

pre- and post-transplantation organoid-derived cells to our primary reference. We noted increased cell subtype specification of both oRG cells and newborn neurons (Extended Data Fig. 14h), suggesting that metabolic stress contributes to specification deficiencies in organoid cells (Supplementary Discussion).

## Conclusions

We have provided a comprehensive molecular characterization of developing human cortical cell types and their preservation in brain organoid models. Using single-cell transcriptomics, we have identified broad cell classes and types, as well as fine-grained subtypes such as outer radial glia progenitors in primary human samples. Compared to primary tissue, organoids contain a smaller number of cell subtypes and their cells often co-express marker genes, resulting in broad type assignment, such as pan-radial glia or pan-neuron. We have used this dataset to generate pseudoage metrics and provide in-depth analysis of area-specific gene signatures and their developmental trajectories in primary and organoid neurons. Finally, we have identified a role for stress pathway activation in the impaired subtype specification of cortical organoid cell types; the lack of specificity in organoids must be carefully considered when studying developmental processes,

cell-type-specific disease phenotypes or cellular connectivity. In addition, metabolic stress in utero could lead to molecular identity changes, with potential consequences for human brain development. Overall, our compilation of raw and analysed data, paired with visualization of single-cell clustering in a cell browser, provides a valuable resource for better understanding of normal human development and to benchmark the fidelity of in vitro cellular data.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1962-0>.

- Kadoshima, T. et al. Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell-derived neocortex. *Proc. Natl Acad. Sci. USA* **110**, 20284–20289 (2013).
- Lancaster, M. A. et al. Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
- Camp, J. G. et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl Acad. Sci. USA* **112**, 15672–15677 (2015).
- Pollen, A. A. et al. Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756.e717 (2019).
- Velasco, S. et al. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523–527 (2019).
- Camp, J. G. et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).
- Wu, H. et al. Comparative analysis and refinement of human PSC-derived kidney organoid differentiation with single-cell transcriptomics. *Cell Stem Cell* **23**, 869–881.e868 (2018).
- Sloan, S. A. et al. Human astrocyte maturation captured in 3D cerebral cortical spheroids derived from pluripotent stem cells. *Neuron* **95**, 779–790.e776 (2017).
- Amiri, A. et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362**, eaat6720 (2018).
- Mansour, A. A. et al. An in vivo model of functional and vascularized human brain organoids. *Nat. Biotechnol.* **36**, 432–441 (2018).
- Eiraku, M. et al. Self-organized formation of polarized cortical tissues from ESCs and its active manipulation by extrinsic signals. *Cell Stem Cell* **3**, 519–532 (2008).
- Xiang, Y. et al. Fusion of regionally specified hPSC-derived organoids models human brain development and interneuron migration. *Cell Stem Cell* **21**, 383–398.e387 (2017).
- Quadrato, G. et al. Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48–53 (2017).
- Giandomenico, S. L. et al. Cerebral organoids at the air-liquid interface generate diverse nerve tracts with functional output. *Nat. Neurosci.* **22**, 669–679 (2019).
- Marton, R. M. et al. Differentiation and maturation of oligodendrocytes in human three-dimensional neural cultures. *Nat. Neurosci.* **22**, 484–491 (2019).
- Paşca, A. M. et al. Human 3D cellular model of hypoxic brain injury of prematurity. *Nat. Med.* **25**, 784–791 (2019).
- Lui, J. H., Hansen, D. V. & Kriegstein, A. R. Development and evolution of the human neocortex. *Cell* **146**, 18–36 (2011).
- Götz, M. & Huttner, W. B. The cell biology of neurogenesis. *Nat. Rev. Mol. Cell Biol.* **6**, 777–788 (2005).
- Nowakowski, T. J. et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).
- Mayer, C. et al. Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457–462 (2018).
- Pollen, A. A. et al. Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015).
- Nowakowski, T. J., Pollen, A. A., Sandoval-Espinosa, C. & Kriegstein, A. R. Transformation of the radial glia scaffold demarcates two stages of human cerebral cortex development. *Neuron* **91**, 1219–1227 (2016).
- Vaid, S. et al. A novel population of Hopx-dependent basal radial glial cells in the developing mouse neocortex. *Development* **145**, dev169276 (2018).
- Harrington, A. J. et al. MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *eLife* **5**, e20059 (2016).
- Barbosa, A. C. et al. MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. *Proc. Natl Acad. Sci. USA* **105**, 9391–9396 (2008).
- Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
- Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
- Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030.e1016 (2018).
- Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e1022 (2018).
- Cadwell, C. R., Bhaduri, A., Mostajo-Radji, M. A., Keefe, M. G. & Nowakowski, T. J. Development and arealization of the cerebral cortex. *Neuron* **103**, 980–1004 (2019).
- Simi, A. & Studer, M. Developmental genetic programs and activity-dependent mechanisms instruct neocortical area mapping. *Curr. Opin. Neurobiol.* **53**, 96–102 (2018).
- Yoshida, A. & Tani, K. Phosphoglycerate kinase abnormalities: functional, structural and genomic aspects. *Biomed. Biochim. Acta* **42**, S263–S267 (1983).
- Izumi, K. et al. *ARCN1* mutations cause a recognizable craniofacial syndrome due to COP1-mediated transport defects. *Am. J. Hum. Genet.* **99**, 451–459 (2016).
- Gee, H. Y., Noh, S. H., Tang, B. L., Kim, K. H. & Lee, M. G. Rescue of  $\Delta F508$ -CFTR trafficking via a GRASP-dependent unconventional secretion pathway. *Cell* **146**, 746–760 (2011).
- Kim, J. et al. Monomerization and ER relocalization of GRASP is a requisite for unconventional secretion of CFTR. *Traffic* **17**, 733–753 (2016).
- Laguesse, S. et al. A dynamic unfolded protein response contributes to the control of cortical neurogenesis. *Dev. Cell* **35**, 553–567 (2015).
- Tseng, K. Y. et al. MANF is essential for neurite extension and neuronal migration in the developing cortex. *eNeuro* **4**, ENEURO.0214-17.2017 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

### PSC expansion culture

The human induced PSC lines H28126 (Gilad Laboratory, University of Chicago), 13234 and WTC10 (Conklin Laboratory, Gladstone Institutes), which were previously authenticated<sup>4</sup>, and the embryonic stem cell line H1 (WiCell, authenticated at source), were expanded on matrigel-coated six-well plates. Cells tested negative for mycoplasma. Stem cells were thawed in StemFlex Pro Medium (Gibco) containing 10  $\mu$ M Rock inhibitor Y-27632. Medium was changed every other day and lines were passaged when colonies reached about 70% confluency. Stem cells were passaged using PBS-EDTA and residual cells were manually lifted with cell lifters (Fisher). All lines used for this study were between passage 25 and 40.

### Cortical organoid differentiation protocols

Cortical organoids were differentiated using three directed differentiation protocols. In brief, PSC lines were expanded and dissociated to single cells using Accutase. After dissociation, cells were reconstituted in neural induction medium at 10,000 cells per well in a 96-well v-bottom low-adhesion plate. After 18 days, organoids from all protocols were transferred from 96-well to 6-well low-adhesion plates and moved onto an orbital shaker rotating at 90 rpm. Throughout the culture duration organoids were fed every other day. Organoids were collected for immunohistochemistry and scRNA-seq after 3, 5, 8 or 10 weeks of culture.

For the least directed differentiation protocol<sup>1</sup>, GMEM-based induction medium included 20% knockout serum replacer (KSR), 1 $\times$  non-essential amino acids, 0.11 mg/ml sodium pyruvate, 1 $\times$  penicillin–streptomycin, 0.1 mM  $\beta$ -mercaptoethanol, 5  $\mu$ M SB431542 and 3  $\mu$ M IWR1-endo. Medium was supplemented with 20  $\mu$ M Rock inhibitor Y-27632 for the first 6 days. After 18 days, the medium was changed to DMEM/F12 medium containing 1 $\times$  glutamax, 1 $\times$  N2, 1 $\times$  CD lipid concentrate and 1 $\times$  penicillin–streptomycin. After 35 days, organoids were moved into DMEM/F12-based medium containing 10% FBS, 5  $\mu$ g/ml heparin, 1 $\times$  N2, 1 $\times$  CD lipid concentrate and 0.5% matrigel (BD). After 70 days, the medium was additionally supplemented with 1 $\times$  B27 and the matrigel concentration was increased to 1%.

In the directed differentiation protocol<sup>4,8</sup>, the induction medium consisted of GMEM including 20% KSR, 1 $\times$  non-essential amino acids, 0.11 mg/ml sodium pyruvate, 1 $\times$  penicillin–streptomycin and 0.1 mM  $\beta$ -mercaptoethanol supplemented with 5  $\mu$ M SB431542, 3  $\mu$ M IWR1-endo and 2  $\mu$ M dorsomorphin. From days 9 to 25 small molecules were removed and the induction medium was instead supplemented with 10 ng/ml EGF and 10 ng/ml GGF. After 25 days the medium was changed to DMEM/F12 medium containing 1 $\times$  glutamax, 1 $\times$  N2, 1 $\times$  CD lipid concentrate and 1 $\times$  penicillin–streptomycin. After 35 days, organoids were moved into DMEM/F12-based medium containing 10% FBS, 5  $\mu$ g/ml heparin, 1 $\times$  N2, 1 $\times$  CD lipid concentrate and 0.5% matrigel (BD). After 70 days, the medium was additionally supplemented with 1 $\times$  B27 and the matrigel concentration was increased to 1%.

The most directed<sup>12</sup> protocol used a DMEM/F12-based induction medium containing 15% KSR, 1 $\times$  MEM-NEAA, 1 $\times$  glutamax, 100  $\mu$ M B-ME, 100 nM LDN-193189, 10  $\mu$ M SB431542, and 2  $\mu$ M XAV939. For the first 2 days, the medium was supplemented with 50  $\mu$ M Rock inhibitor Y-27632 and 5% heat-inactivated FBS. After 10 days, organoids were moved into neuronal differentiation medium consisting of equal parts DMEM/F12 and neurobasal medium containing 0.5% N2, 1% B27 without vitamin A, 1% glutamax, 0.5% MEM-NEAA, 0.025% human insulin solution, 50  $\mu$ M B-ME and 1% penicillin–streptomycin. After 18 days, organoids were maintained in maturation medium containing equal parts DMEM/F12 and neurobasal medium with 0.5% N2, 1% B27, 1% glutamax, 0.5% NEAA,

0.025% human insulin solution, 50  $\mu$ M B-ME, 20 ng/ml BDNF, 200  $\mu$ M cAMP and 200  $\mu$ M ascorbic acid.

### Immunohistochemistry

Cortical organoids and primary human cortical tissue samples were collected, fixed in 4% PFA, washed with 1 $\times$  PBS and submerged in 30% sucrose in 1 $\times$  PBS until saturated. Samples were embedded in cryomolds containing 50% OCT (Tissue-tek) and 50% of 30% sucrose in 1 $\times$  PBS and frozen at  $-80^{\circ}\text{C}$ . Primary samples were sectioned at 20  $\mu$ M and organoids at 16  $\mu$ M onto glass slides. Antigen retrieval was performed on tissue sections using a citrate-based antigen retrieval solution at 100 $\times$  (Vector Labs) which was boiled to  $95^{\circ}\text{C}$  and added to slides for 20 min. After antigen retrieval, slides were briefly washed with PBS and blocked with PBS containing 5% donkey serum, 2% gelatin and 0.1% Triton for 30 min. Primary antibodies were incubated in blocking buffer on slides at  $4^{\circ}\text{C}$  overnight, washed with PBS containing 0.1% Triton three times and then incubated with AlexaFluor secondary antibodies (Thermo Fisher) at room temperature for 2 h. Primary antibodies included mouse: SOX2 (Santa Cruz, 1:500, sc-365823), HOPX (Santa Cruz, 1:250, sc-398703), SATB2 (Abcam, 1:250, ab51502), AUTS2 (Abcam, 1:100, ab243036), human nuclei (Millipore, 1:500, MAB1281); rabbit: HOPX (Proteintech, 1:500, 11419-1-AP), GORASP2 (Proteintech, 1:50, 10598-1-AP), ARCN1 (Proteintech, 1:50, 23843-1-AP), PGK1 (Thermo Fisher 1:50, PA5-13863), PTPRZ1 (Atlas, 1:250, HPA015103), NR2F1 (Novus, 1:100, NBPI-31259); rat: CTIP2 (Abcam, 1:500, ab18465); sheep: EOMES (R&D, 1:200, AF6166); guinea pig: NEUN (Millipore, 1:500, ABN90); and chicken: GFP (Aves, 1:500, GFP-1020).

### Primary sample collection

All primary tissue was obtained and processed as approved by the UCSF Human Gamete, Embryo and Stem Cell Research Committee (GESCR, approval 10-05113). All experiments were performed in accordance with protocol guidelines. Informed consent was obtained before sample collection for the use of all tissue samples within this study. First and second trimester human cortex tissue was collected from elective pregnancy termination specimens from San Francisco General Hospital and the Human Developmental Biology Resource (HDBR). Tissue was collected only with previous patient consent for research and in strict observation of legal and institutional ethical regulations.

### Dissociation

Primary human cortical samples were dissociated using papain (Worthington) containing DNase. Samples were grossly chopped and then placed in 1 ml papain and incubated at  $37^{\circ}\text{C}$  for 15 min. Samples were inverted three times and incubation continued for another 15 min. Next, samples were triturated by manually pipetting with a glass pasteur pipette approximately ten times. Dissociated cells were spun down at 300g for 5 min and papain removed.

### 10 $\times$ capture and sequencing

Single-cell capture from live cells was performed following the 10 $\times$  v2 Chromium manufacturer's instructions for both primary and organoid samples. For primary samples, each sample was its own batch. For organoid samples, batch is indicated in the metadata annotation in Supplementary Table 1. In each case, 10,000 cells were targeted for capture and 12 cycles of amplification for each of the cDNA amplification and library amplification were performed. Libraries were sequenced as per manufacturer recommendation on a NovaSeq S2 flow cell.

### Clustering

We first explored the cell-type identities of primary and organoid samples using Louvain-Jaccard clustering<sup>19,38</sup>. Prior to clustering, batch correction was performed in a similar way to previous approaches<sup>39</sup>. In brief, each set of cells within a batch was normalized to the



# Article

highest expressing gene, making the range of expression from 0 to 1. These values were multiplied by the average number of counts within the batch. These normalized datasets were piped into Seurat v.2<sup>40</sup>, in which cells with fewer than 500 genes per cell or more than 10% of reads aligning to mitochondrial genes were discarded. Normalized counts matrices were  $\log_2$ -transformed, and variable genes were calculated using default Seurat parameters. Data were scaled in the space of these variables, and the batch was regressed out. Principal component analysis was performed using FastPCA, and significant principal components were identified using a published formula<sup>38</sup>. In the space of these significant principal components, the  $k=10$  nearest neighbours were identified as per the RANN R package. The distances between these neighbours were weighted by their Jaccard distance, and Louvain clustering was performed using the igraph R package. If any clusters contained only one cell, the process was repeated with  $k=11$  and upwards until no clusters contained only one cell. Cluster markers and  $t$ -SNE plots were generated with Seurat package default parameters.

## Cell-type annotations

Primary cell-type annotations of clusters were performed by comparison to previously annotated cell types, and when a repository of substantial matching was not available, a combination of literature-based annotation of layer or maturation stage identity was used. The genes used to annotate each cluster are highlighted in Supplementary Table 2. When a cluster was substantially enriched based upon an age or an areal metadata property, this empirical observation was used to inform the annotation. Organoid cell types were first annotated by their similarity to primary cell clusters; if the correspondence was at or above 0.4 and only one primary cell type had such a high correspondence, the primary cell type was applied to the organoid cluster. If the correspondence was between 0.2 and 0.4 and included only one similarity, that cell type was used to identify the organoid cell type unless there was an obvious discrepancy in top marker gene expression between the two clusters. If no correlation was above 0.2, literature annotations or unknown identities were assigned. If an organoid cluster correlated equally well (within 10%) with multiple primary subtypes of the same or similar cell type, 'pan' identity was assigned. Low-quality cell types for all analyses were assigned when markers were dominated (>60%) by mitochondrial genes, ribosomal genes, or pseudogenes. Occasionally, an intersection of these approaches was used for organoid clusters, as indicated in Supplementary Table 3.

## Correlation analysis

Correlation analysis was generated in the space of marker genes. For each marker gene, a specificity score was calculated. This score equalled the 'enrichment' ( $\log_2$ (fold change) of the marker compared to other clusters) and the 'specificity' (the percentage of the relevant cluster expressing the marker divided by the percentage of other clusters expressing the marker). These two values were multiplied by one another to obtain the final score, and this was represented across all marker genes for each sample in box-and-whisker plots. A matrix of all markers across all clusters was created for each individual dataset; if a marker was not expressed at all in a certain cluster, it was marked as 0. If a value was divided by 0 to calculate the score, the score was placed as a dummy score at 1,500. Matrices between comparisons were correlated in the space of overlapping marker gene space using Pearson's correlations.

## Co-clustering analysis

Each of the five batch correction methods was performed with default parameters. For each analysis, the same 20,000 cell subset of each organoid and primary cells was used because most of the algorithms were too computationally intensive to perform on the full dataset.

## Linear mixed models

VariancePartition<sup>41</sup> was used for linear mixed model analysis. Analysis was performed in a randomized subset of 50,000 genes in the space of expressed genes across the metadata properties noted in Extended Data Fig. 9. Age was used as a continuous variable and all other variables were assigned as discrete.

## WGCNA and maturation analysis

WGCNA networks were calculated as previously described<sup>19</sup> in 10,000 randomly chosen primary radial glia cells and in parallel from 10,000 randomly chosen organoid radial glia cells. These networks were applied to the remaining primary and organoid cells using the ModuleEigengene function from the WGCNA R package. Pseudoage was calculated by taking networks that correlated highly to age in the 10,000 cell subset and combining their genes into a single gene set. Principal component analysis was performed in this gene space in the full space of radial glia and the loading of the first principal component dictated the pseudoage. This analysis was performed reciprocally.

## Area signatures

Area signatures were obtained by performing pairwise differential expression between each of the seven cortical areas and the six remaining areas. Differential expression across all of the areas was combined, with a count of how many times a gene was differentially expressed in an area from each of the pairwise comparison. While combining the lists, the enrichment and specificity were averaged across all six analyses and multiplied by the number of times the gene appeared as a marker for an area of interest. This value, the 'area specificity score', was compared across all areas. For any genes that were considered markers of multiple areas, the area with the highest area specificity score was allocated to the gene as a marker, thus making all area markers unique to one area alone. This is how some areas have a higher percentage of cells assigned to another area other than their area of origin, and enables cleaner comparison of areal pattern emergence. Each set of area marker genes were designated as a network, and the correlation of each cell to this area was calculated by applyModules and calculating a module eigengene. After assignment, to normalize unequal module eigengene distributions, within a dataset the module eigengenes were normalized by area and the assigned area for a cell was the area for which that cell had the highest module eigengene.

## PSA-NCAM protocol and viral infection of primary cells

Primary cortical samples were grossly dissected to isolate the ventricular and subventricular zones excluding cortical plate neurons. Dissociated cells were enriched for neural progenitor cells using a PSA-NCAM antibody and the MACs magnetic sorting kit (Miltenyi Biotec). In brief, dissociated cells were incubated with the PSA-NCAM antibody for 30 min at room temperature, washed with PBS containing 0.1% BSA and added to an equilibrated magnetic L column. Cells positive for antibodies bind to the column and negative cells are collected in the elute. The negatively sorted cells were pelleted at 300g for 10 mins and supernatant removed. Negatively sorted samples were infected with a CMV::GFP adenovirus (Vector Biolabs), which preferentially labels progenitors, at 37 °C for 15 min. Cells were spun down for 5 min at 300g and reconstituted in 500  $\mu$ l medium. Cells were counted and 50,000 primary cells isolated for transplantation in 100  $\mu$ l medium.

## Primary cell transplantation into organoids

Week-10 or -15 organoids made from the 13234 induced PSC line using the least directed differentiation protocol were placed at the air-liquid interface on Millicell (Millipore) inserts to limit movement. Fifty thousand primary cortical cells were reconstituted in medium containing DMEM/F12, 5% FBS, 1 $\times$  N2 (Thermo Fisher), 1 $\times$  B27 (Thermo Fisher), 1 $\times$  penicillin-streptomycin (Thermo Fisher) and CD lipid concentrate

(Thermo Fisher). Cells were slowly pipetted on top of the semi-dry organoids and left to integrate into the organoids for 30 min at 37 °C. Afterwards, organoids were gently lifted off of inserts by increasing medium volume by 3 ml. After 2 days, organoids were transferred to a new 6-well plate without inserts and the medium supplemented with 1% GF-reduced matrigel and 1× amphotericin B.

#### FACS purification of GFP<sup>+</sup> cells

Cells were dissociated into a single-cell suspension as described above. Cell suspensions were triturated and placed on top of 4 ml 22% Percoll. Tubes containing Percoll and cell suspension were spun at 500g for 10 min without break. The supernatant was discarded, and the cell pellet resuspended in HBSS with BSA and glucose. Cells were sorted using a Becton Dickinson FACSaria using 13 psi pressure and a 100-µm nozzle aperture. All FACS gates were set using unlabelled cells. Data were analysed post hoc for enrichment percentages with FlowJo software.

#### Maintenance of transplanted organoids

Organoids were maintained in 6-well low-adhesion plates in DMEM/12 with glutamax (Thermo) medium containing 10% FBS (Hyclone), 1% GF-reduced matrigel (Corning), 1× N2 (Thermo Fisher), 1× B27 (Thermo Fisher), 1× CD lipid concentrate (Thermo Fisher), 5 µg/ml heparin, 1× penicillin–streptomycin, and 1× amphotericin B (Gibco). The medium was changed every other day for the duration of culture. Transplants were collected for immunohistochemistry at weeks 1, 2.5, 4 and 6 post-transplant. At week 2.5, paired organoid samples were FACS-sorted for GFP<sup>+</sup> cells and captured cells were collected for single-cell RNA sequencing.

#### Organoid cell transplantation into mice

Mouse experiments were approved by UCSF Institutional Animal Care and Use Committee (IAUCUC) protocol AN178775-01 and performed in accordance with relevant institutional guidelines. Organoids from the H28126 and 13234 induced PSC lines were differentiated using the least directed protocol described above. Week-7/8 organoids were dissociated and labelled with a CMV::GFP adenovirus for 30 min at 37 °C. Cells were pelleted and immediately transplanted into postnatal NSG mice (NOD.Cg-Prkdc<sup>scid</sup> Il2rg<sup>tm1Wjl</sup>/Sz), stock No:005557) at four days of age (P4). Using a stereotaxic rig, either the PFC or V1 of the left hemisphere was localized and 10,000 cells were transplanted into the cortex at each injection site. Two transplantations or injections, 0.5 mm apart, were made per mouse within the same cortical area of 10,000 cells each. Mice were allowed to develop for either 2 or 5 weeks post-transplantation before being euthanized. A total of 13 mice were used (7 male, 6 female), and no statistical method was used to determine this sample size. Mice were euthanized, and the brain was extracted and grossly dissected using GFP expression to visualize relevant areas for collection. Tissue with GFP expression was dissociated for 30 min at 37 °C using papain, manually triturated, centrifuged and resuspended in HBSS. Cells were sorted for GFP using the FACS strategy described previously. After FACS isolation, GFP<sup>+</sup> cells were used for scRNA-seq using the 10× V2 platform. Mice from the same experiment were collected in parallel and perfused with 10 ml PBS followed by 10 ml 4% PFA before the brains were extracted. Mouse brains were further fixed overnight at 4 °C, washed in 1× PBS three times for 30 min each, and rocked overnight at 4 °C in 30% sucrose in PBS. Brains were embedded in 50/50 mix of 30% sucrose and OCT before being cryosectioned. Mouse studies were not randomized or blinded prior to analysis.

#### Organotypic slice culture

Primary cortical tissue was maintained in CO<sub>2</sub>-bubbled artificial cerebral spinal fluid until being embedded in a 3% low melt agarose gel. Embedded tissue was live-sectioned at 300 µm using a vibratome (Leica) and plated on Millicell (Millipore) inserts in a 6-well tissue culture plate. Slices were cultured at the air–liquid interface in medium containing 32% Hanks BSS, 60% BME, 5% FBS, 1% glucose, 1% N2 and

1% penicillin–streptomycin–glutamine. Slices were maintained for 7 days in culture at 37 °C and the medium was changed every third day.

#### Primary cortical aggregates

Primary human cortical samples from gestational weeks 14 and 15 were gross-dissected at the outer subventricular zone, removing the cortical plate. Samples were dissociated using papain as described previously. Samples were aggregated in 96 v-bottom low-adhesion plates (S bio) containing 20,000 cells per well. They were aggregated in DMEM/F12 with Glutamax (Thermo) based medium containing 10% FBS (Hyclone), 1× N2 (Thermo Fisher), 1× CD lipid concentrate (Thermo Fisher), 5 µg/ml Heparin, 1× penicillin–streptomycin and 1× amphotericin B (Gibco). Medium was supplemented with 20 µM Rock inhibitor for the first week. After 2 weeks, aggregates were transferred to 6-well low-adhesion plates and medium was supplemented with 1% matrigel and 1× B27.

#### Dissociated primary cell culture

Dissociated primary cortical cells were reconstituted in DMEM/F12-based medium containing 1× N2 (Thermo Fisher), 1× B27 (Thermo Fisher), 1× penicillin–streptomycin (Thermo Fisher) and 1× sodium pyruvate (Thermo Fisher). Cells were plated at one million cells per ml in 12-well matrigel-coated tissue culture plates.

#### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### Data availability

Single-cell RNA sequencing data have been deposited in dbGAP for accession 'A cellular resolution census of the developing human brain' and in GSE132672. An interactive browser of single-cell data and raw and processed count matrices can be found at the UCSC cell browser website: <https://organoidreportcard.cells.ucsc.edu>. Source Data for Figs. 1–5 and Extended Data Figs. 1–14 are available online. Remaining source data can be retrieved directly from the single-cell data available in public repositories or from the UCSC cell browser website.

38. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323.e1330 (2016).
39. Peng, Y. R. et al. Molecular classification and comparative taxonomics of foveal and peripheral cells in primate retina. *Cell* **176**, 1222–1237.e1222 (2019).
40. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
41. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
42. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
43. Velmeshev, D. et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685–689 (2019).

**Acknowledgements** We thank Q. Bi, S. Wang, W. Walantus, C. Villareal, A. Alvarez-Buylla, C. Kim, O. Meyerson and members of the Kriegstein laboratory for resources, technical help and helpful discussions. This study was supported by NIH award U01MH114825 to A.R.K., and F32NS103266 and K99NS111731 to A.B., as well as by the California Institute for Regenerative Medicine (CIRM) through the CIRM Center of Excellence in Stem Cell Genomics (GC1R-06673-C to A.R.K.).

**Author contributions** A.B., M.G.A., A.A.P., T.J.N. and A.R.K. designed the study and analysis. Experiments were performed by M.G.A., W.M.L., Diane Jung, A.B., D.S., D.A., Dana Jung, G.S. and J.S. Data analysis was performed by A.B., M.G.A. and M.H. The study was supervised by A.B., M.G.A. and A.R.K. This manuscript was prepared by A.B. and M.G.A. with input from all authors.

**Competing interests** The authors declare no competing interests.

#### Additional information

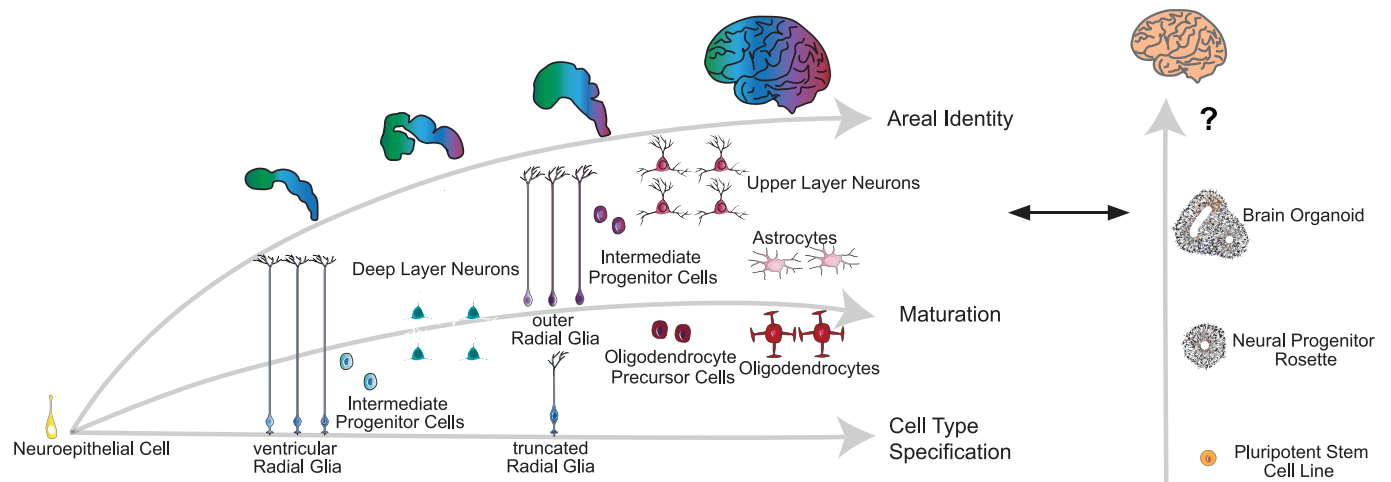
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1962-0>.

**Correspondence** and **requests for materials** should be addressed to A.R.K.

**Peer review information** Nature thanks Andrew Adey, Flora Vaccarino and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

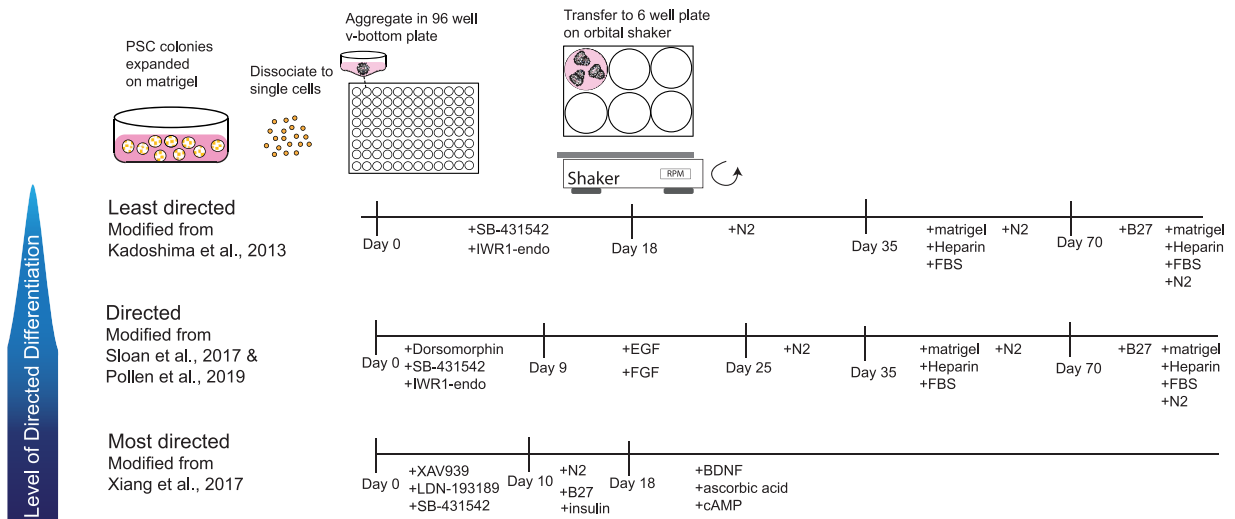
# A Comparing Axes of Biological Variation of Development Between Primary Human Cortex and Cortical Organoids



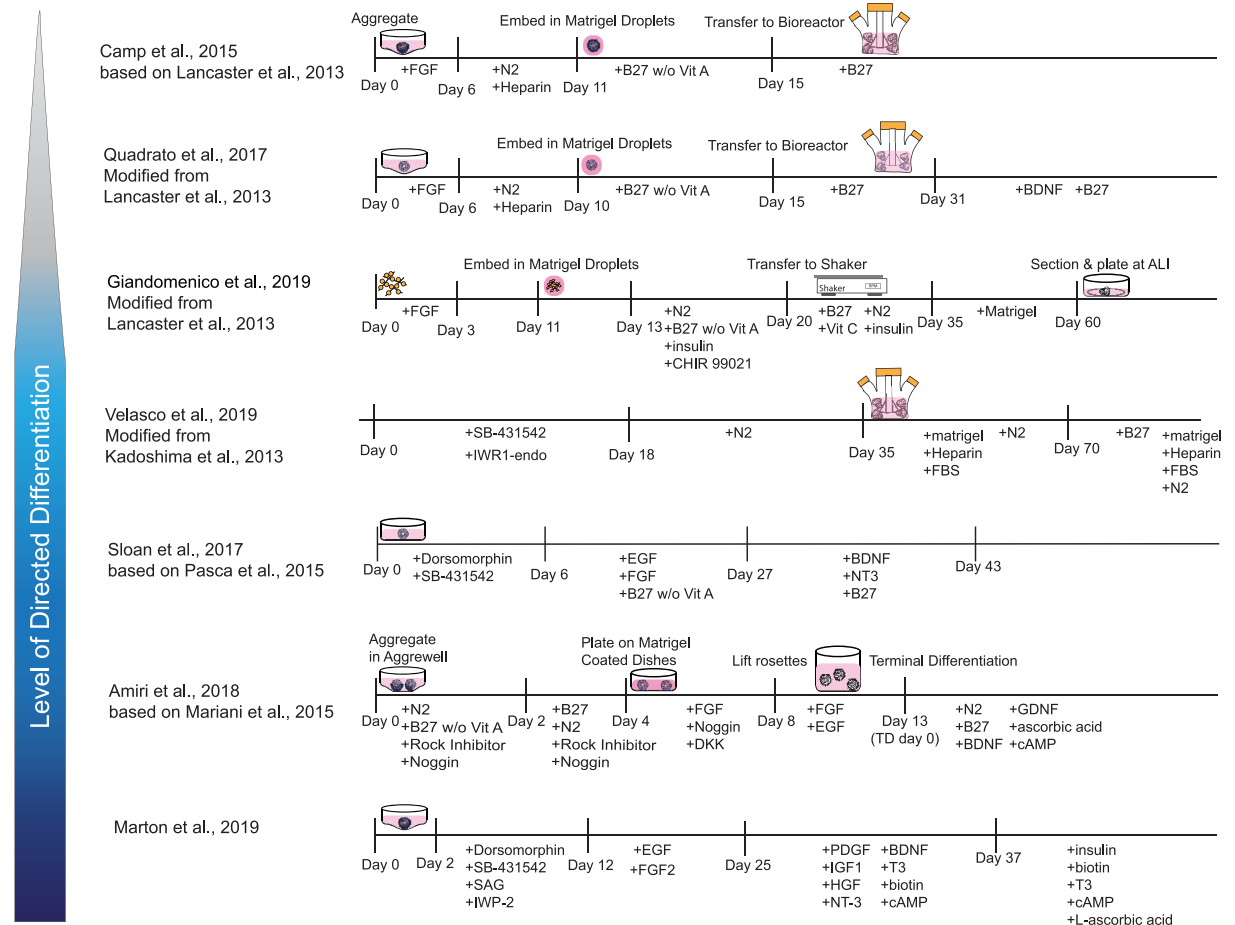
**Extended Data Fig. 1 | Schematic of human cortex and human organoid development. a,** Schematic of normal brain developmental trajectories queried in this study and their comparison to organoid models. Normal cortical development requires the emergence of a diversity of progenitor cell types from a seemingly uniform neuroepithelium. Through a sequence of cell-

type specification and maturation, progenitor cells undergo neurogenesis and gliogenesis to generate the cellular diversity of the cortex. Areal identities are specified during this process and comprise a core property of developing neurons.

## A Organoid Protocols Utilized for Single Cell RNA Sequencing



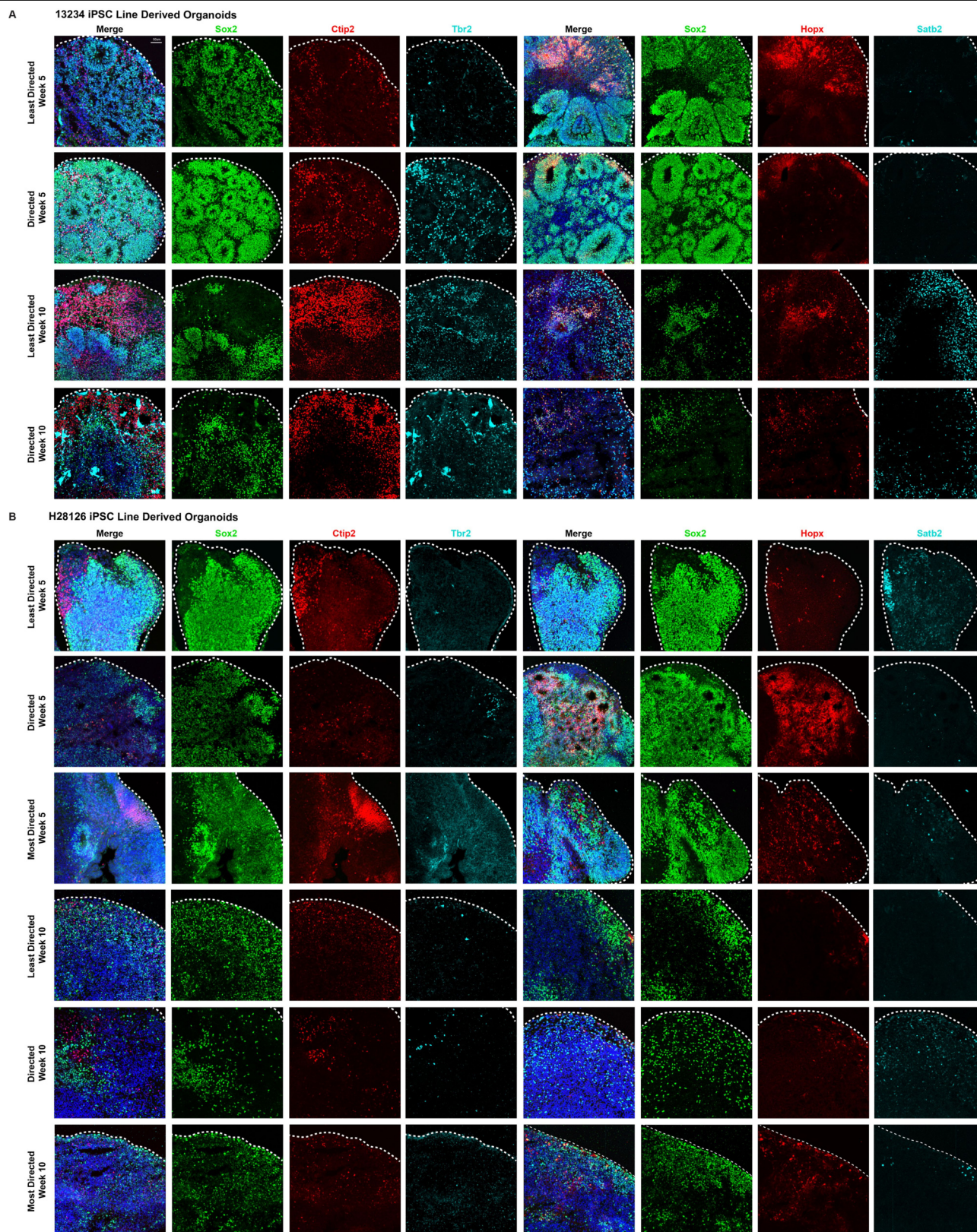
## B Publicly Available Organoid Data Sets Compared in this Analysis



**Extended Data Fig. 2 | Brain and cortical organoid generation protocols. a,** Cortical organoid protocols using different levels of directed differentiation were evaluated using scRNA-seq and immunohistochemistry. Stem cells were expanded on matrigel, dissociated to single cells, and re-aggregated in v-bottom low-adhesion plates. Small molecules were used to promote

forebrain induction and after 18 days cells were moved to 6-well plates on an orbital shaker. Organoids were maintained in culture and collected from weeks 3 to 24. **b,** Protocol schematics for other methods used to differentiate whole brain and cortical organoids, which have published single-cell data. Publicly available data were used for comparative analyses with our collected data.

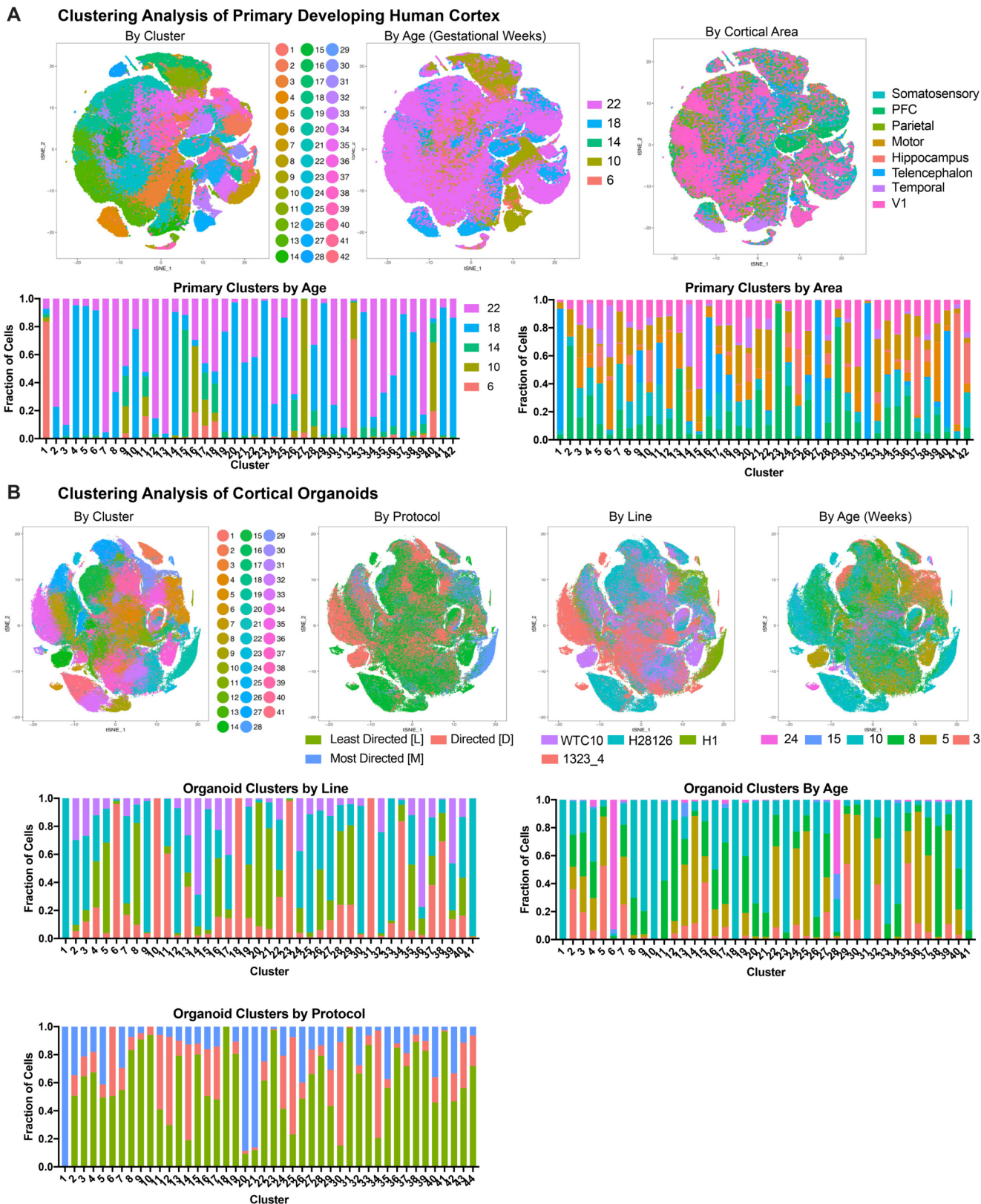




**Extended Data Fig. 3 | Comparison of broad cell types across differentiation protocols.** **a**, Organoids derived from the 13234 induced PSC line underwent the least and directed differentiation protocols, were collected at weeks 5 and 10, and were processed for immunohistochemistry. Organoids from both protocols were stained with SOX2 to mark progenitors, HOPX to identify outer radial glia and TBR2 to label intermediate progenitor cells. Cultures were also stained with CTIP2 to mark deep layer neurons and SATB2 to identify upper

layer neurons. At week 5 all progenitor subtypes were present, and by week 10 both deep and upper layer neurons were detected. **b**, Organoids from the H28126 induced PSC line were differentiated using the least, directed and most directed protocols. All progenitor types marked by SOX2, HOPX and TBR2 and CTIP2<sup>+</sup> and SATB2<sup>+</sup> neuronal populations were present by week 5. Expression of all markers decreased by week 10. Organoid staining validation of broad cell types was repeated independently three times.

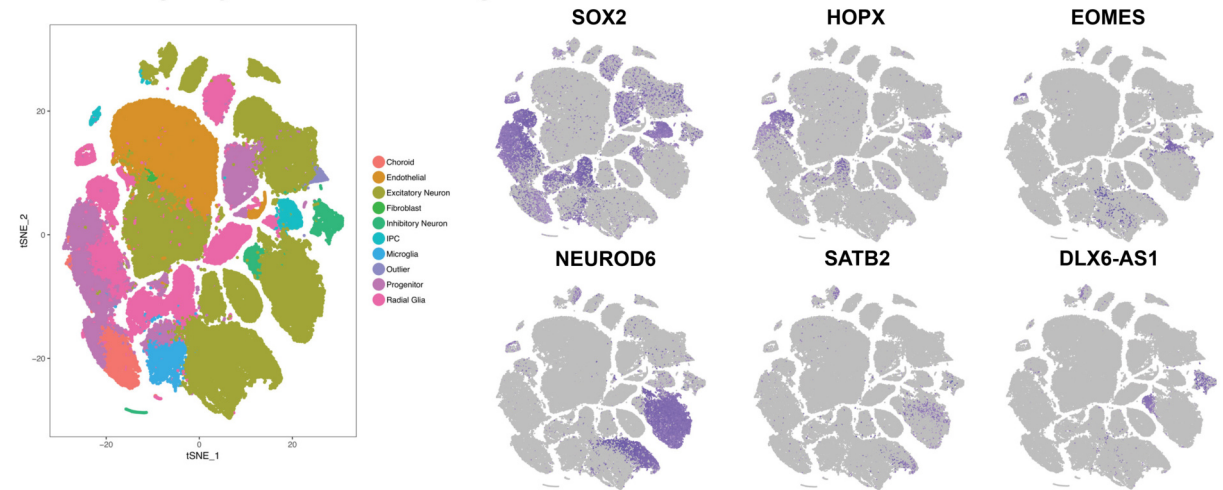




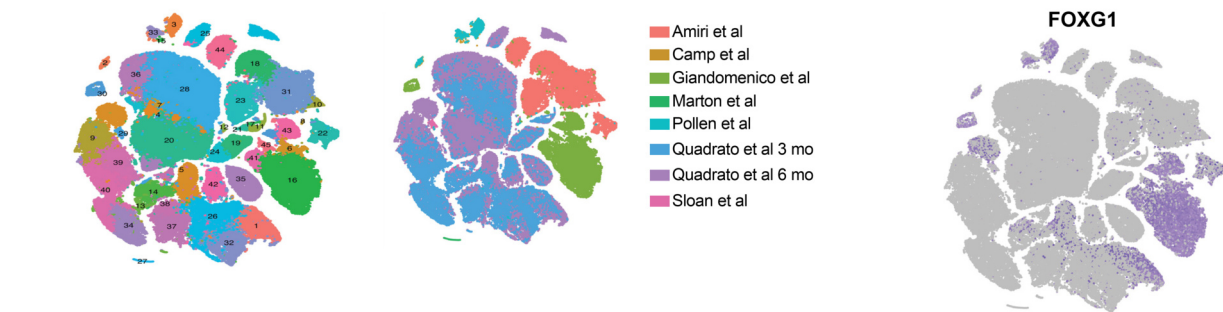
**Extended Data Fig. 4 | Single-cell comparison of cell types across samples.**  
**a**, *t*-SNE plots depicting the single-cell analysis of primary cortical cells as coloured by cluster, age of sample and cortical area. Stacked histograms showing composition of each cluster for these metadata properties are also

included. **b**, *t*-SNE plots depicting the single-cell analysis of cortical organoid cells as coloured by cluster, protocol, pluripotent stem cell line (induced PSC or human embryonic stem cell) and age of sample. Stacked histograms showing composition of each cluster for these metadata properties are also included.

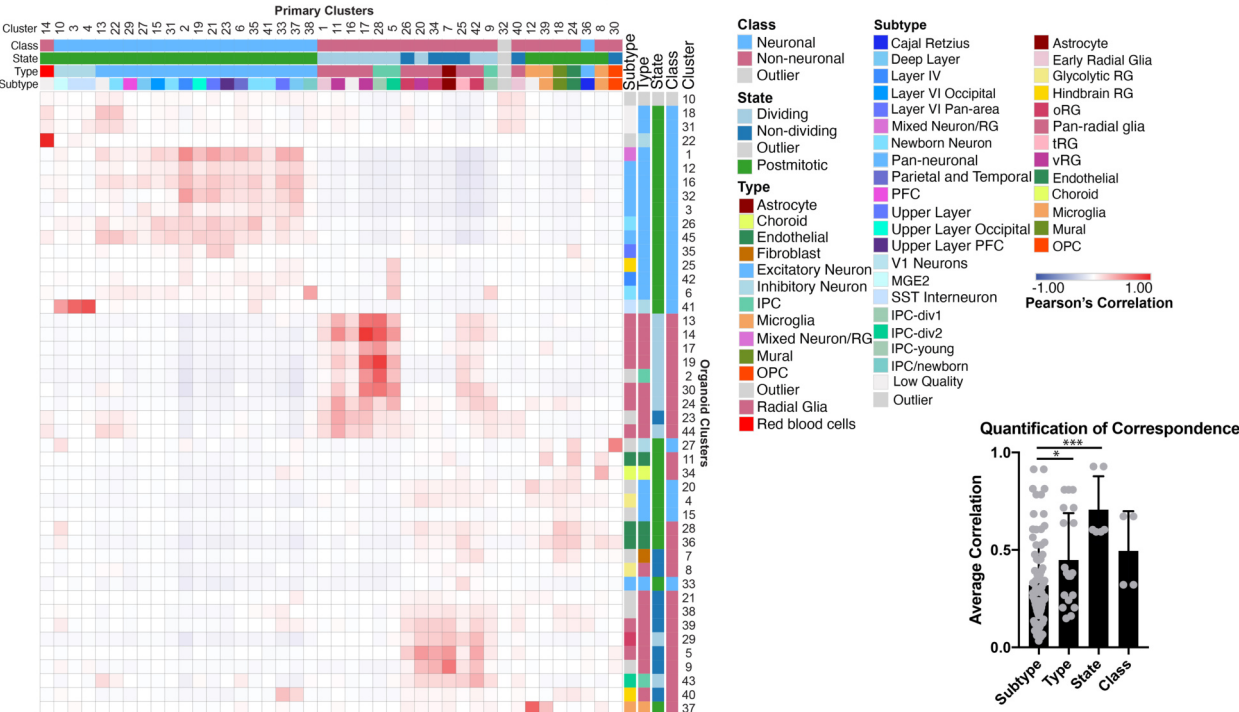
A Clustering Analysis of Published Brain Organoid Datasets



B Composition of Published Brain Organoid Datasets

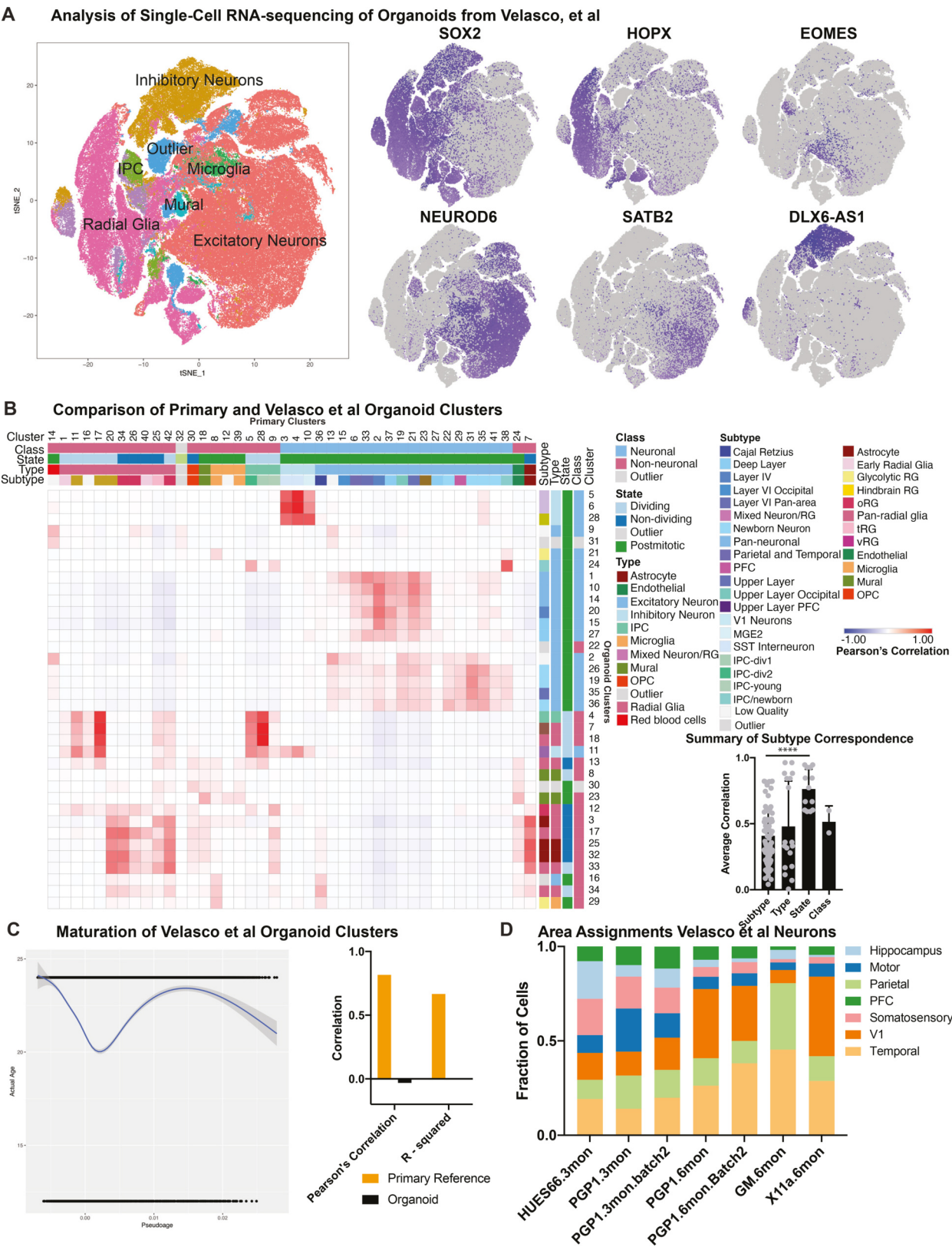


C Comparison of Published Organoid Datasets to Primary Human Brain



**Extended Data Fig. 5 | Single-cell comparison of cell types across published datasets.** **a**, Re-analysis of published single-cell sequencing in organoid samples. *t*-SNE plot is coloured by cell-type designation, and the feature plots depict the same cell populations as presented in Fig. 1. **b**, *t*-SNE plots depicting the single-cell analysis of published organoid cells as coloured by cluster, protocol (including paper of origin) and FOXG1 expression. **c**, Recapitulation of the heat map in Fig. 2, using published organoid clusters from above and comparing to primary reference dataset from this paper. Quantification of

correspondence shows the quantitative correlation from the best match in the heat map for each category of class, state, type and subtype, averaged across all clusters (primary:  $n=189,409$  cells from five individuals collected independently; published organoid data:  $n=109,813$  cells from 7 datasets collected independently by different scientific groups; two-sided Welch's *t*-test evaluating mean + s.d.; subtype versus type,  $*P=0.0193$ ; subtype versus state,  $***P=0.00017$ ).



Extended Data Fig. 6 | See next page for caption.

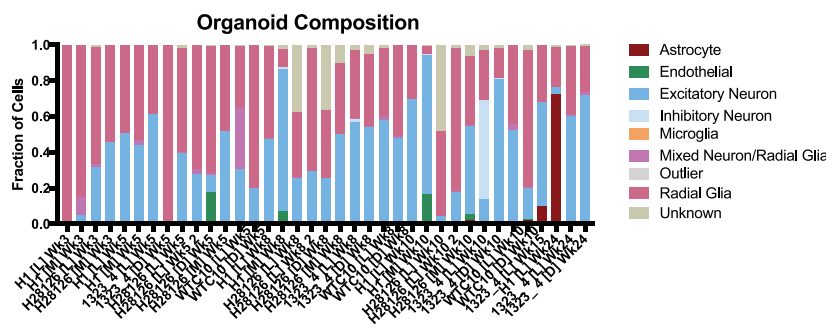
# Article

**Extended Data Fig. 6 | Single-cell comparison of cell types across published datasets.** **a**, Re-analysis of published single-cell sequencing<sup>5</sup>, in which a reproducible cortical organoid protocol was presented. The *t*-SNE plot is coloured by cell-type designation, and the feature plots depict the same cell populations as presented in Fig. 1. **b**, Recapitulation of the heat map in Fig. 2, using published organoid clusters<sup>5</sup> and comparing to the primary reference dataset from this paper. Quantification of correspondence shows the quantitative correlation from the best match in the heat map for each category of class, state, type and subtype, averaged across all clusters (primary:  $n = 189,409$  cells from five individuals collected independently; organoid data<sup>5</sup>:

$n = 166,241$  cells from an independently collected dataset; two-sided Welch's test was used to evaluate mean + s.d.; subtype versus state \*\*\*\* $P = 1.8 \times 10^{-7}$ ). **c**, Pseudoage analysis of published organoids<sup>5</sup> mirrors the organoids in this study with low correspondence between pseudoage and actual age. Pseudoage calculation is indicated by the graph line and shading represents the geometric density standard error of the regression. **d**, Area identity was assigned for all excitatory neurons from ref. <sup>5</sup> and each organoid consisted of heterogeneous areal identities, consistent with the observations in the organoids from this study.



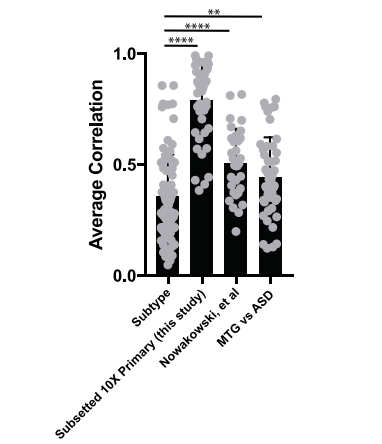
**A Reproducibility and Composition of Organoids**



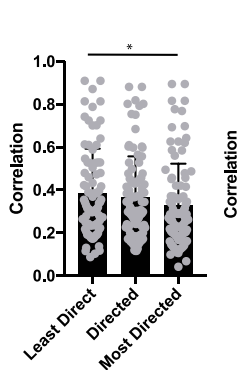
**FOXP1**



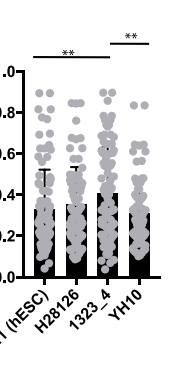
**B Subtype Correspondence Summary**



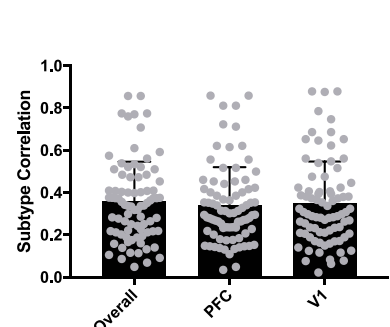
**C Subtypes By Protocol**



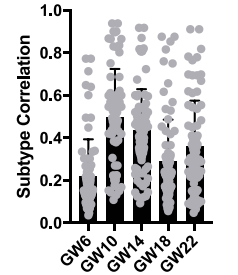
**Subtypes By Line**



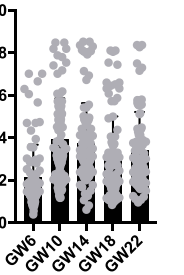
**D Subtype By Key Areas**



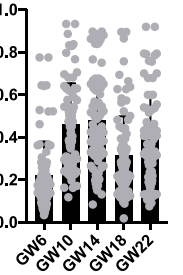
**E Week 3 Organoids**



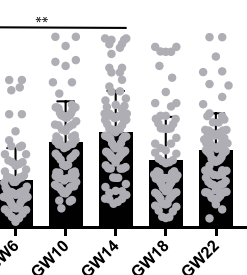
**Week 5 Organoids**



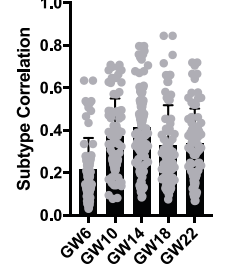
**Week 8 Organoids**



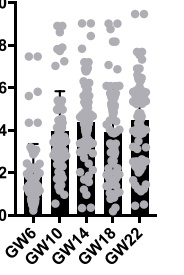
**Subtype By All Ages**



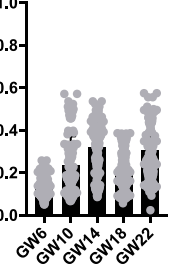
**Week 10 Organoids**



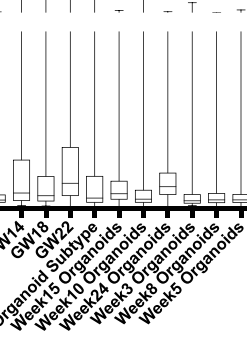
**Week 15 Organoids**



**Week 24 Organoids**



**Specificity By Age**



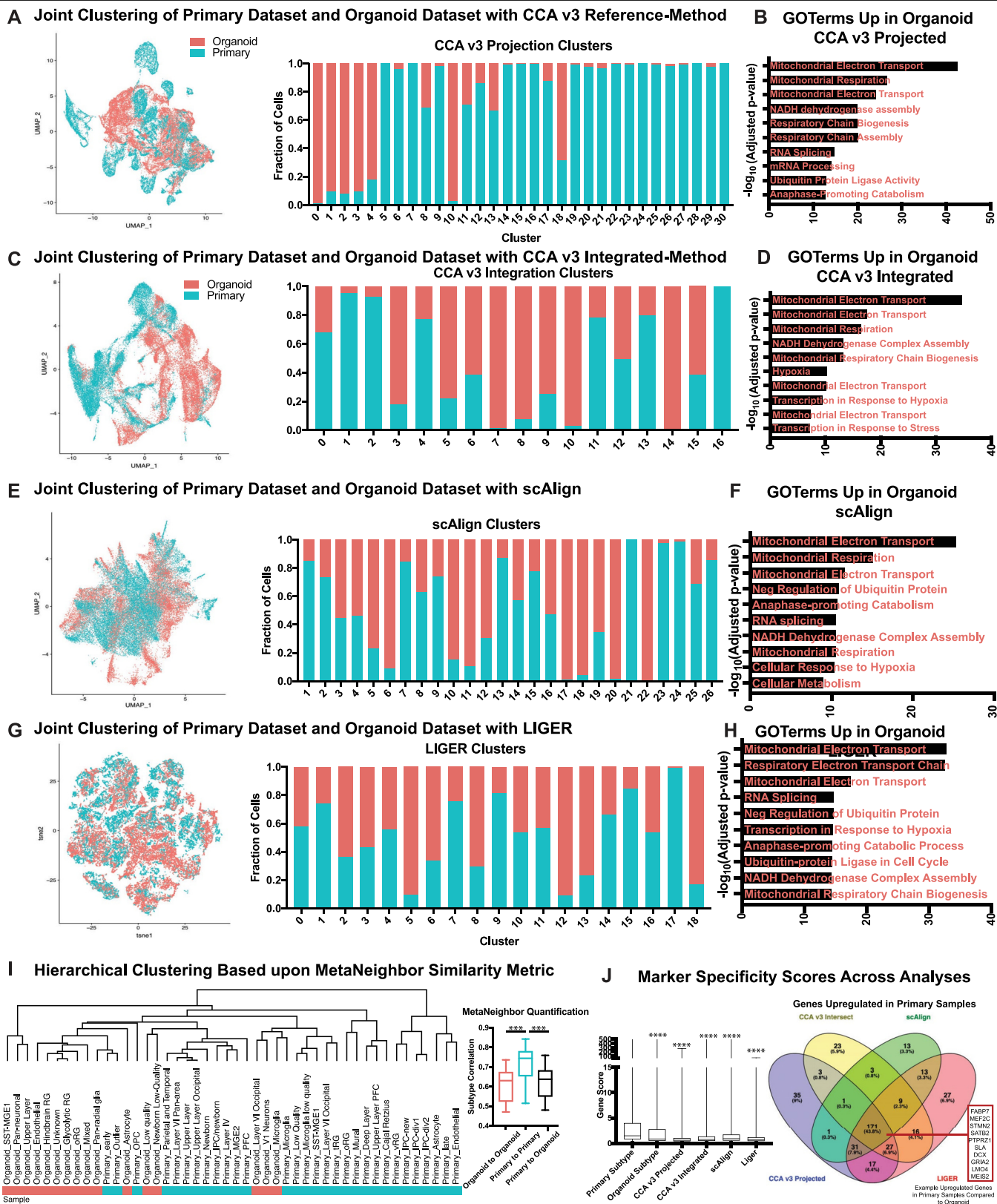
**Extended Data Fig. 7** | See next page for caption.



## Extended Data Fig. 7 | Analysis of subtype correlation across metadata properties.

**a.** Composition of each organoid by cell-type designation. FOXG1 expression across all organoid samples is plotted by feature on the right. **b.** Comparison of organoid subtype as determined by this study versus three control analyses. Graphically, the column indicates subtype correspondence; error bar, s.d. The first analysis was performed by halving the primary dataset randomly and without overlap and then comparing the subclusters from the two datasets. This age- and method-matched analysis shows that primary variation is significantly lower than the variation between organoids and primary cells, as indicated by the significantly higher subtype correlation between primary datasets (organoids:  $n = 242,349$  cells collected from 37 organoids from 4 biologically independent samples from 4 independent experiments; primary data: 189,409 cells from 5 biologically independent samples from 5 experiments; \*\*\*\* $P = 2.0 \times 10^{-24}$ , two-sided Welch's  $t$ -test). A similar analysis was performed comparing the primary data from this study to data collected by microfluidic approaches<sup>19</sup>. Although the ages, capture method and number of cells varied greatly, subtype correlation between the published primary data and the data in this study is significantly higher than the subtype similarity between organoids and primary samples<sup>19</sup> ( $n = 4,261$  cells from 48 biologically independent samples across more than 35 independent experiments, \*\*\*\* $P = 2.0 \times 10^{-5}$ ). We additionally performed this analysis between two published datasets for cells from adult humans, comparing middle temporal gyrus<sup>42</sup> (MTG,  $n = 15,928$  cells) from an older adult with distinct brain regions from young adults in the control samples of a study on autism spectrum disorder<sup>43</sup> (ASD,  $n = 104,559$ ). Despite differences across ages and individuals, who could be expected to have unique cortical gene-expression profiles based upon sensory experience, the distinct cortical regions isolated and the different capture methods, the subtype correlation between these two primary datasets is significantly higher than the correlation between organoid cells and primary cells (\*\* $P = 0.0076$ ). **c.** Subtype correlation

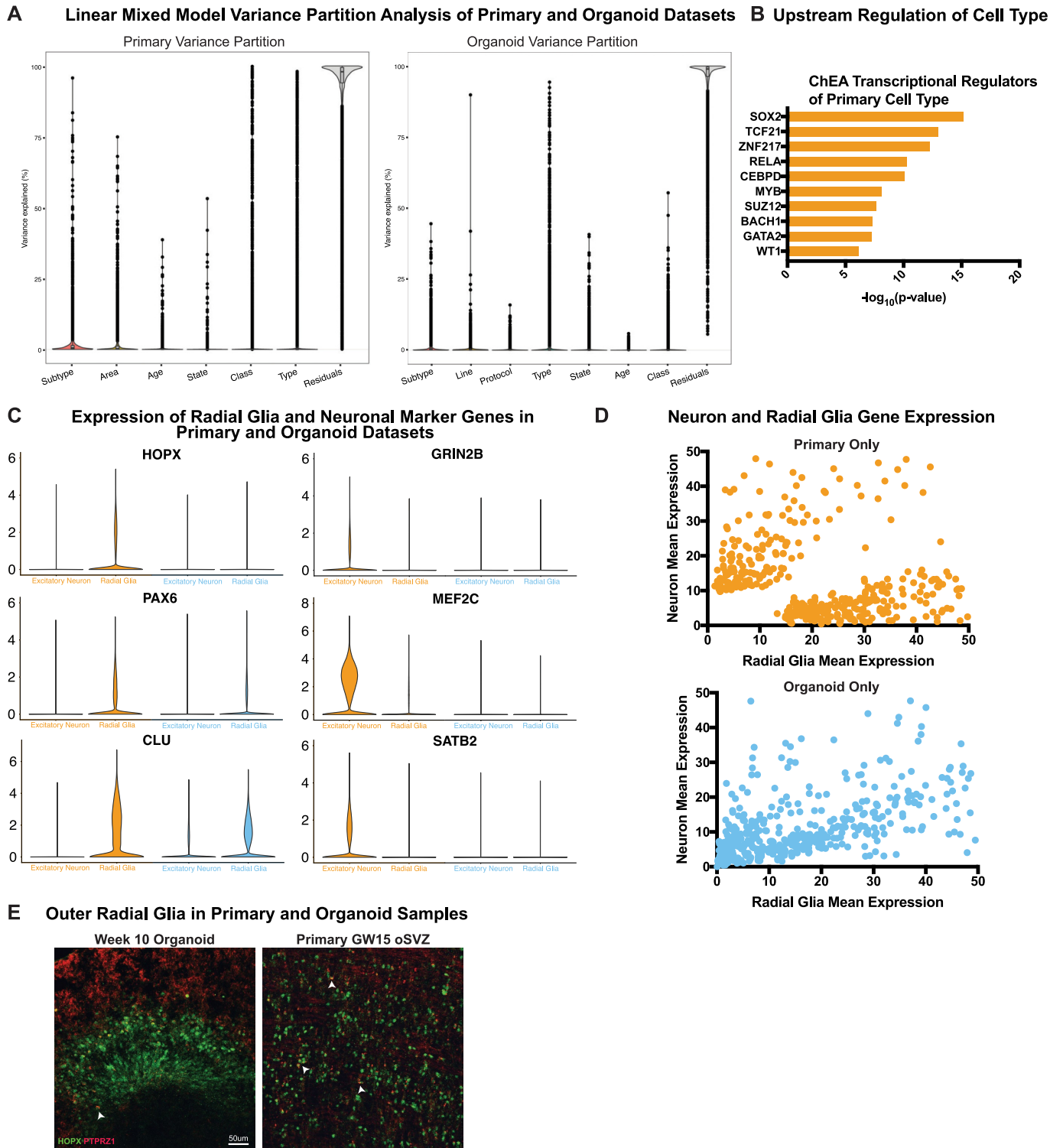
as calculated and shown in Fig. 2, broken down by protocol and pluripotent line, in which bars indicate subtype correlation and error bars show s.d. The least directed protocol was significantly better at recapitulating cell subtype than the most directed protocol (\* $P = 0.0483$ , two-sided Welch's  $t$ -test), consistent with recent findings<sup>5</sup>. We also observed that the induced PSC line 1323\_4 generated significantly more similar subtypes to primary samples than WTC10 or H1 (\*\* $P = 0.0013$  and  $0.0089$ , respectively). **d.** Clustering and subtype analysis was performed between all organoids and primary PFC samples and primary V1 individually. Subtype correlation did not change regardless of the area to which organoids were compared. 'Overall' refers to the subtype correlation observed when comparing all organoids cells to all primary cells and is shown for comparison. Histogram bars show subtype correlation and error bars show s.d. ( $n = 242,349$  cells from 37 organoids across 4 independent experiments). **e.** Subtype correlation analysis was performed across all organoid stages ( $n =$  week 3: 38,417 cells, week 5: 26,787 cells, week 8: 11,023 cells, week 10: 50,550 cells, week 15: 2,722 cells, week 24: 4,506 cells from 4 independent experiments) and all primary ages ( $n =$  GW6: 5,970 cells, GW10: 7,194 cells, GW14: 14,435 cells, GW18: 78,157 cells, GW22: 83,653 cells from 5 independent experiments). Histogram bars show subtype correlation and error bars show s.d. Week-3 organoids are more similar to younger primary stages, and week-15 organoids are most similar to older primary ages. Other ages correspond similarly well to the primary stages of peak neurogenesis (GW10–24), and altogether the organoids are most significantly similar to GW14 (\*\* $P = 0.0015$ , two-sided Welch's  $t$ -test). 'Overall' refers to the subtype correlation observed when comparing all organoids cells to all primary cells and is shown for comparison. The last histogram shows the average gene score of each sample and error bars show s.d. Younger primary samples and organoids have a relatively lower gene score related to their marker specificity; this specificity increases substantially over time in primary cells but less so in organoid cells.



**Extended Data Fig. 8** | See next page for caption.

**Extended Data Fig. 8 | Co-clustering of primary and organoid single-cell datasets with CCA, scAlign, LIGER and MetaNeighbour.** **a**, Canonical correlation analysis from Seurat v3 was performed using reference-based integration. For this analysis, 20,000 cells were randomly subsetting from both the primary and organoid datasets and their counts matrices were merged. The primary samples were designated as the reference, and using CCA the organoid cells were projected into that reference space. A UMAP plot of the intersection is shown. The stacked histogram shows the relative contributions of each sample to each cluster. Most clusters were primarily one dataset or the other, validating the observations of limited primary subtype recapitulation in organoids. **b**, For the clusters with at least 20% contribution from both primary and organoid cells, differential expression was performed across all of these clusters jointly using a two-sided Wilcoxon rank-sum test. The full differential expression is presented in Supplementary Table 5, but genes upregulated in organoid cells were examined with Enrichr pathway analysis, and a summary of the top Gene Ontology terms is presented (organoid:  $n = 20,000$  cells from 37 organoids across 4 independent experiments; primary:  $n = 20,000$  cells from 5 individuals across 5 independent experiments). **c**, Canonical correlation analysis from Seurat v3 was performed using the integration-based method. For this analysis, 20,000 cells were randomly subsetting from both the primary and organoid datasets and their counts matrices were merged. A UMAP plot of the intersection is shown. The stacked histogram shows the relative contributions of each sample to each cluster. Most clusters were primarily one dataset or the other, validating the observations of limited primary subtype recapitulation in organoids. **d**, For the clusters with at least 20% contribution from both primary and organoid cells, differential expression was performed across all of these clusters jointly using a two-sided Wilcoxon rank-sum test. The full differential expression is presented in Supplementary Table 5, but genes upregulated in organoid cells were examined with Enrichr pathway analysis, and a summary of the top Gene Ontology terms is presented (organoid:  $n = 20,000$  cells from 37 organoids across 4 independent experiments; primary:  $n = 20,000$  cells from 5 individuals across 5 independent experiments). **e**, scAlign was performed for integration of datasets. For this analysis, 20,000 cells were randomly subsetting from both the primary and organoid datasets and their counts matrices were merged. A UMAP plot of the intersection is shown. The stacked histogram shows the relative contributions of each sample to each cluster. Many clusters were primarily one dataset or the other, validating the observations of limited primary subtype recapitulation in organoids. **f**, For the clusters with at least 20% contribution from both primary and organoid cells, differential expression was performed across all of these clusters jointly using a two-sided Wilcoxon rank-sum test. The full differential expression is presented in Supplementary Table 5, but genes upregulated in organoid cells were examined with Enrichr pathway analysis, and a summary of the top Gene Ontology terms is presented (organoid:  $n = 20,000$  cells from

37 organoids across 4 independent experiments; primary:  $n = 20,000$  cells from 5 individuals across 5 independent experiments). **g**, LIGER was performed for integration of datasets. For this analysis, 20,000 cells were randomly subsetting from both the primary and organoid datasets and their counts matrices were merged. A UMAP plot of the intersection is shown. The stacked histogram shows the relative contributions of each sample to each cluster. Although the clusters were well mixed, they had very diffuse marker gene expression suggesting key biological drivers of variation were obscured by the analysis. **h**, For the clusters with at least 20% contribution from both primary and organoid cells, differential expression was performed across all of these clusters jointly using a two-sided Wilcoxon rank-sum test. The full differential expression is presented in Supplementary Table 5, but genes upregulated in organoid cells were examined with Enrichr pathway analysis, and a summary of the top Gene Ontology terms is presented (organoid:  $n = 20,000$  cells from 37 organoids across 4 independent experiments; primary:  $n = 20,000$  cells from 5 individuals across 5 independent experiments). **i**, MetaNeighbour was performed using unsupervised analysis to compare the clusters from primary and organoid samples. MetaNeighbour uses cell-cell similarity scores based upon neighbour voting and AUROC calculations to quantify the similarities between cells. These pairwise values were used as an input to hierarchical clustering, and almost entirely segregated primary clusters from organoid clusters. Box-and-whiskers plot shows quantification of the similarities within organoid and primary datasets versus the comparison of the two showed that the primary alone comparisons were significantly higher (organoid to organoid:  $***P = 0.00078$ ; primary to organoid:  $***P = 0.00036$ , two-sided Welch's  $t$ -test) (organoid:  $n = 20,000$  cells from 37 organoids across 4 independent experiments; primary:  $n = 20,000$  cells from 5 individuals across 5 independent experiments). The bars show range of subtype correlation with middle line indicating the mean and error bars the maximum and minimum. These results further validate our observations that there are important distinctions between the organoid and primary subtypes. **j**, The gene score for each of the four integration methods is presented, and all are significantly lower than for primary clustering alone (organoid subtype:  $****P = 5.3 \times 10^{-38}$ ; CCA v3 Projected:  $****P = 5.5 \times 10^{-94}$ ; CCA v3 Integrated:  $****P = 2.8 \times 10^{-24}$ ; scAlign:  $****P = 2.1 \times 10^{-23}$ ; LIGER:  $****P = 2.9 \times 10^{-94}$ , two-sided Welch's  $t$ -test). The one method that substantially integrated the samples (LIGER) had the lowest gene score. Box-and-whisker plot shows mean score and error bars show maximum and minimum ( $n = 242,349$  cells from 37 organoids across 4 independent experiments). The differentially expressed genes that were upregulated in primary samples from all four analyses are intersected. A substantial number of these genes were found in all four datasets, and these genes included examples that we identified from other methods in this study, including *PTPRZ1*, *MEF2C* and *SATB2*, validating the accuracy of our analytical methods and our main findings.

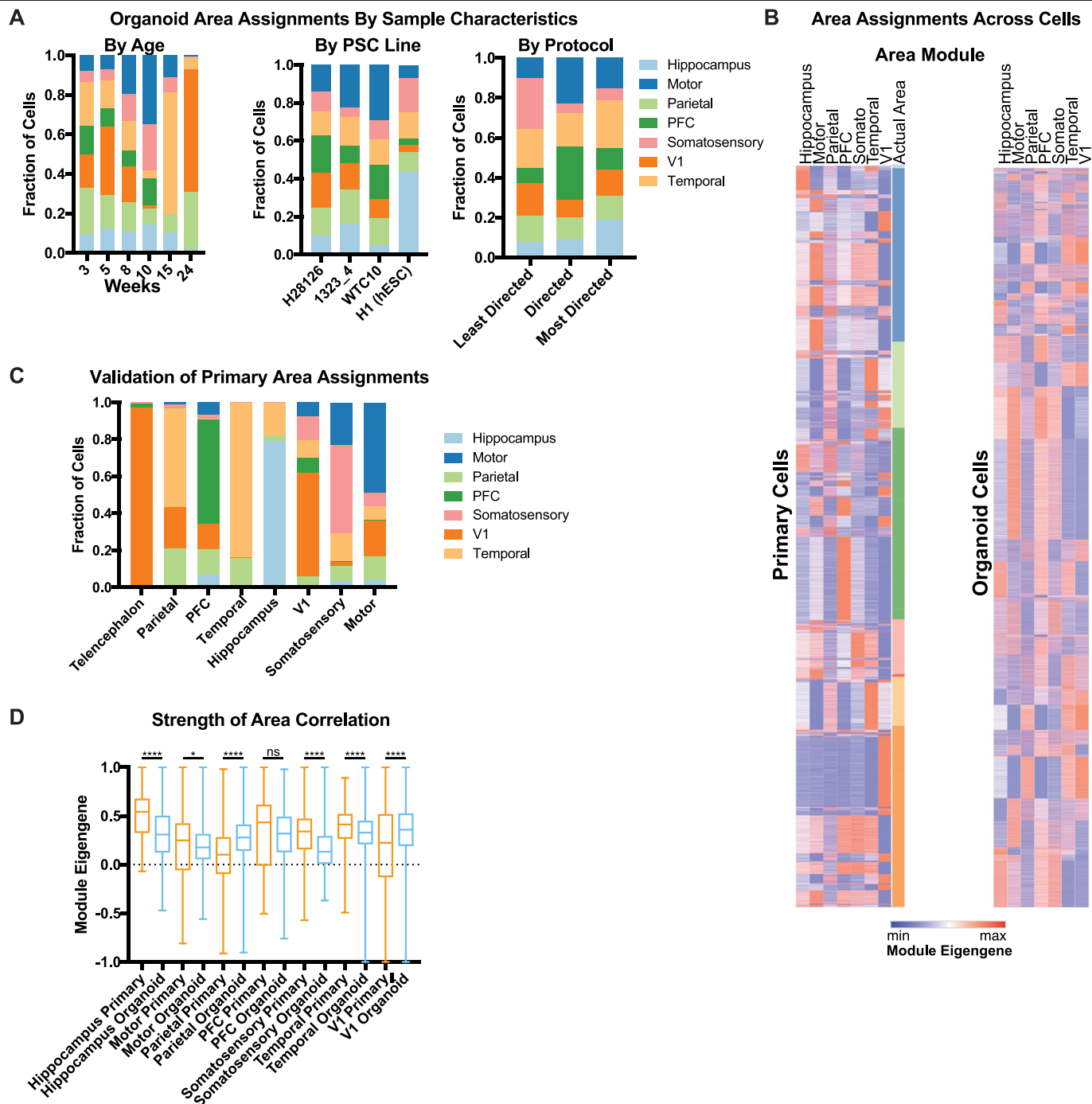


**Extended Data Fig. 9 | Comparing cell-type specification in primary and organoid samples.** **a**, Variance partition was run on both primary and organoid datasets across the metadata properties shown. Each dot represents a gene and the amount of variance of that gene explained by the relevant metadata property. **b**, ChEA analysis of type genes identified in primary cortical samples. The x-axis shows the  $-\log_{10}(\text{adjusted } P)$  of the transcription factors indicated; results obtained from Enrichr datasets included a variety of experimental systems but have been shortened for ease of reading to the relevant transcription factor ( $n = 189,409$  cells from 5 biologically independent samples; two-sided Wilcoxon rank-sum test). Type genes in organoid samples were not unified for significant transcription factor regulation. **c**, Violin plots of radial glia and neuron markers in primary (orange) and organoid (blue) radial

glia and neurons in which width of coloured section indicates distribution of expression of each data point within a sample. In some cases, organoids show expression of multiple markers, lower expression of key markers, or similar expression to that seen in primary samples (organoids:  $n = 242,349$  cells from 37 organoids across 4 independent experiments; primary:  $n = 189,409$  cells from 5 biologically independent samples from 5 independent experiments). **d**, Dot plots from Fig. 2 shown with one colour only to avoid dot overlap. **e**, Lower-magnification images of PTPRZ1 and HOPX overlap as shown in Fig. 2c show domains of overlapping expression in the primary oSVZ and distinct domains of expression in the organoid ventricular zone. Validation stains were repeated independently three times.

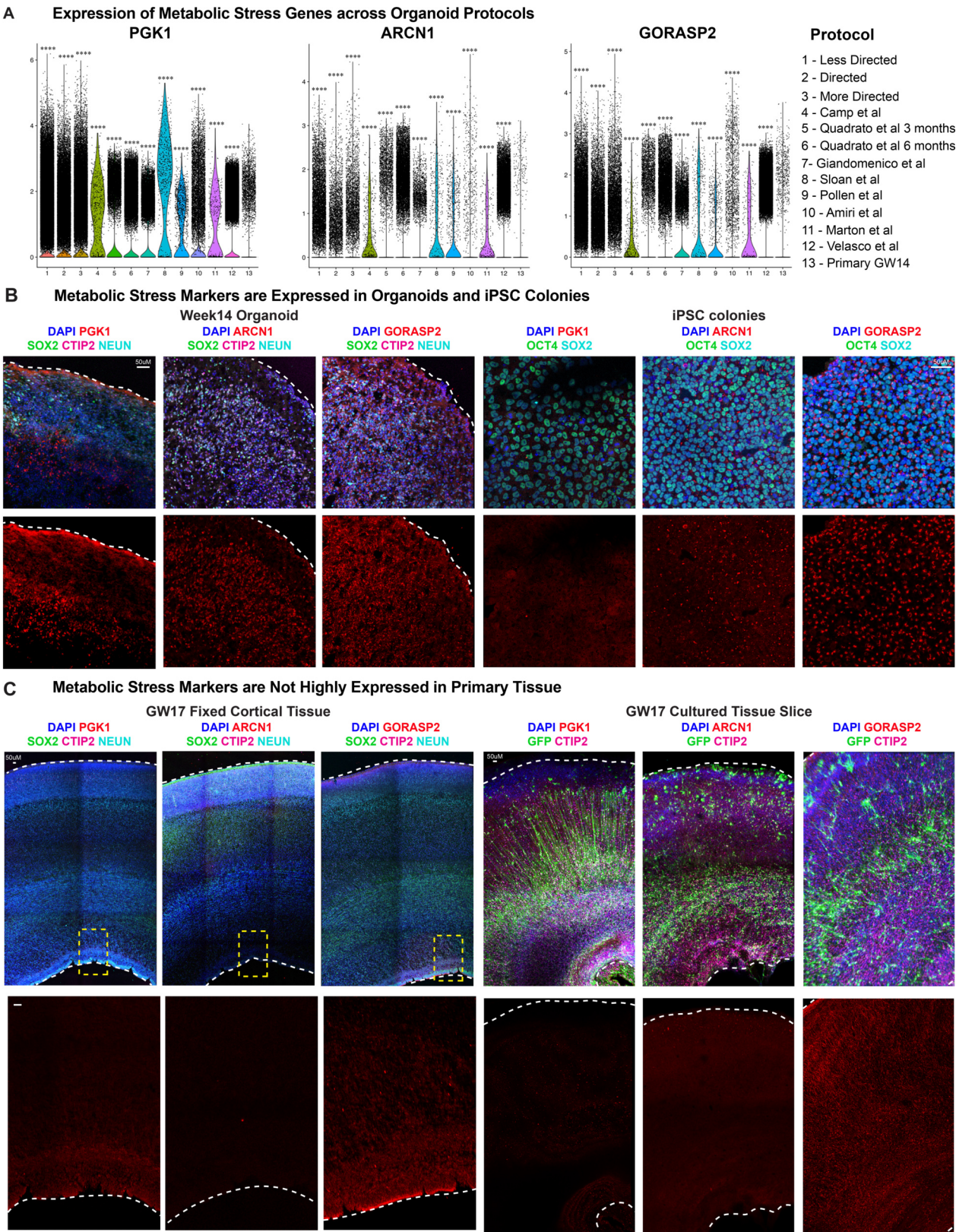






**Extended Data Fig. 11 | Areal identification.** **a**, Organoid areal assignments by age, line and protocol indicate heterogeneous areal identity. **b**, Heat maps showing normalized module eigengene signature of each area in primary samples (with known area on the right) and in organoid samples. **c**, Summary of assigned area in primary samples compared to actual area. In many cases, they correspond strongly, and in others there is evidence of lack of distinction. For example, parietal cells still strongly express temporal signatures, suggesting that they have not yet been distinctly specified in primary samples, although this specification does exist in organoids. **d**, Box-and-whisker plot (minimum to

maximum, bar at mean, error bars show s.d.) is the same comparison as shown in Fig. 4c, but across all areas (primary:  $n = 122,958$  excitatory neurons from 5 individuals from 5 independent experiments; organoids:  $n = 97,531$  excitatory neurons from 37 organoids from 4 biologically independent stem cell lines. In some cases there is no significant difference between strength of area signal in primary cells and organoid cells (PFC, NS (not significant),  $P = 0.5373$ ), in other cases either the primary or organoid sample is significantly stronger (motor:  $*P = 0.0148$ ; all other areas:  $****P < 0.0001$ ; Welch's two-sided  $t$ -test).

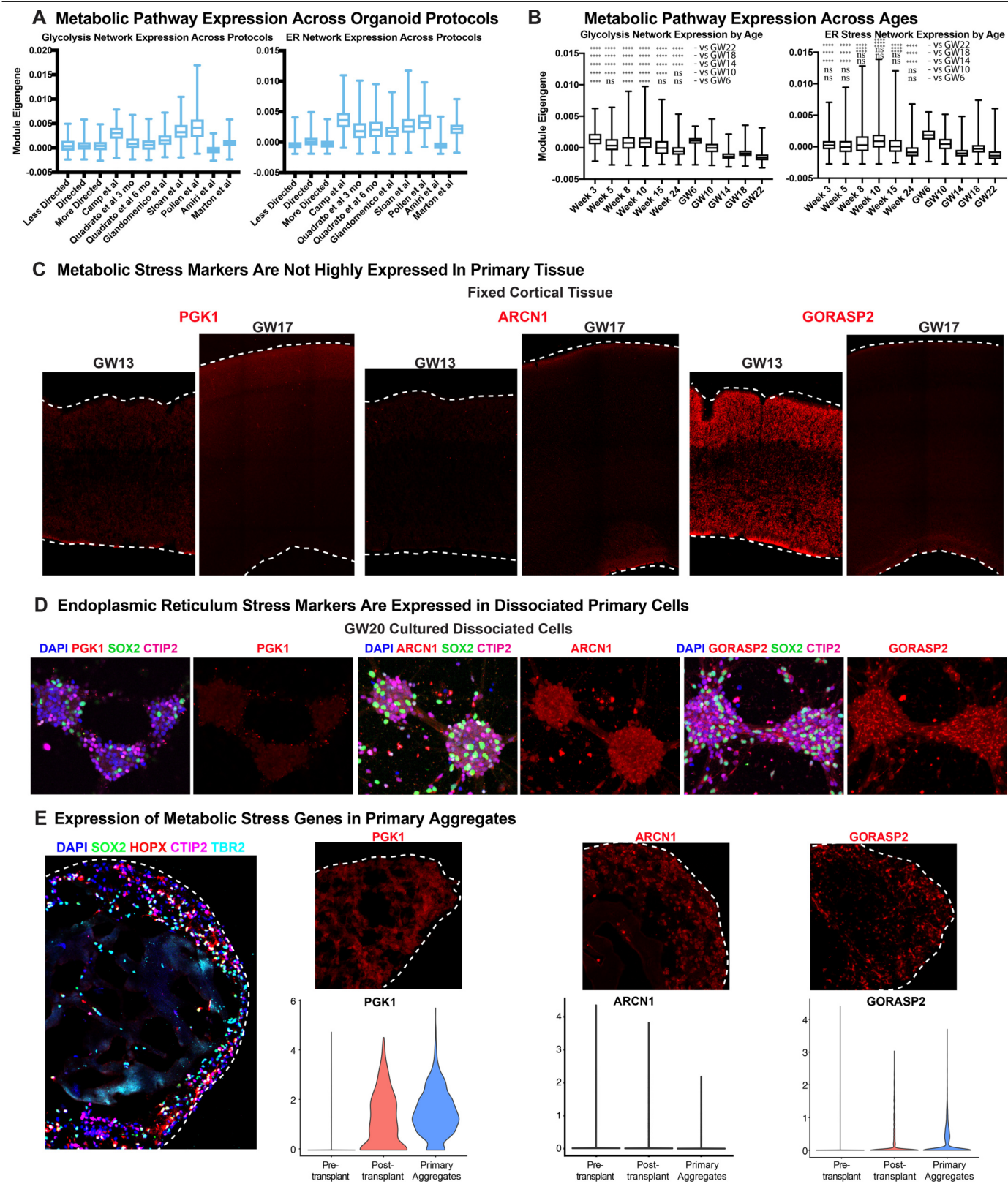


Extended Data Fig. 12 | See next page for caption.

**Extended Data Fig. 12 | Glycolysis and ER stress across culture systems. a,** Markers of metabolic stress are expressed across cortical organoid protocols. Violin plots show both data from our experiments (1–3) and published datasets from other protocols (4–12), which have significantly increased expression of the glycolysis gene *PGK1* and the ER stress genes *ARCN1* and *GORASP2* compared to primary samples ( $n = 5$  individual replicates, GW14 shown). Width of the colored area indicates mean gene-expression level of each dataset and overlaid dots show each individual data point. All protocols have significantly higher expression of these three markers compared to primary samples (\*\*\*\* $P < 0.0001$ , two-sided Student's  $t$ -test). **b,** Single-cell sequencing identified increased expression of genes in organoids, which was validated

across all stages of organoid differentiation evaluated (weeks 3–14). Validation staining experiments were repeated independently three times. Representative images from week-14 organoids differentiated using the least directed differentiation protocol. Colonies of induced PSCs also express the ER stress markers *ARCN1* and *GORASP2* ( $n = 3$  biologically independent samples across 3 experiments). Scale bar, 50  $\mu\text{m}$ . **c,** Primary cortical tissue express glycolysis and ER stress genes at undetectable levels ( $n = 3$  biologically independent samples across 3 experiments). When tissue was cultured for one week, there was no significant increase in cellular stress ( $n = 3$  biologically independent samples across three experiments). Scale bar, 50  $\mu\text{m}$ .





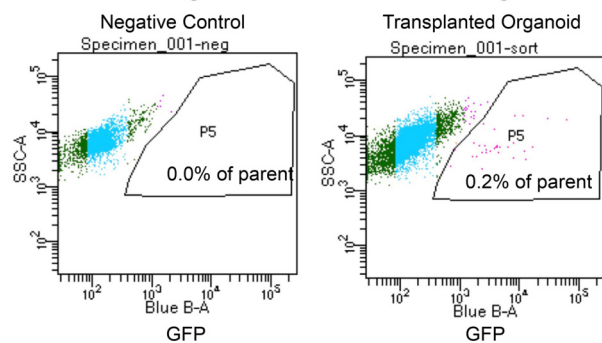
Extended Data Fig.13 | See next page for caption.

**Extended Data Fig. 13 | Glycolysis and ER stress across experimental conditions.** **a**, Metabolic stress network module eigengene expression across all cells is shown in box-and-whisker plots (minimum to maximum, bar at average, error bars show s.d.) across 11 datasets generated either in this manuscript or from publicly available datasets. Data are shown for expressed genes from KEGG pathway glycolysis and ER stress networks. This study:  $n = 242,349$  cells from 37 organoids across 4 independent experiments; published datasets as annotated. **b**, The same box-and-whisker plots are shown for organoids ( $n =$  week 3: 38,417 cells, week 5: 26,787 cells, week 8: 11,023 cells, week 10: 50,550 cells, week 15: 2,722 cells, week 24: 4,506 cells from 4 independent experiments) and all primary ages ( $n =$  GW6: 5,970 cells, GW10: 7,194 cells, GW14: 14,435 cells, GW18: 78,157 cells, GW22: 83,653 cells from 5 independent experiments). ER stress and glycolysis networks decrease over time in primary samples but decrease less in organoids and are significantly higher in most organoid stages than in primary samples. Significance was calculated for each organoid sample with respect to each primary sample, and a one-sided Welch's  $t$ -test was performed (to evaluate whether organoid expression was higher than primary). All comparisons were either not significant (ns) or significant with \*\*\*\* $P < 0.0001$ . **c**, Cellular stress genes are

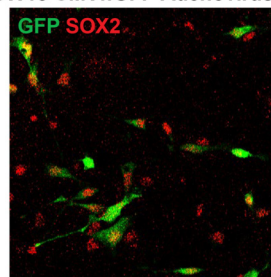
expressed at low levels during human cortical development. GW13 and 17 samples were stained for the glycolysis gene, *PGK1*, and showed little expression at either age. The ER stress gene *ARCNI* had little expression at either age, but there was modest expression of the ER stress gene *GORASP2* at GW13 that decreased by later neurogenesis. Staining validation studies were performed independently four times. **d**, Dissociated primary cells were cultured for one week. Across five independent studies, there was no detectable expression of the glycolysis gene *PGK1*, but the ER stress genes *ARCNI* and *GORASP2* showed significantly increased expression. **e**, Immunostaining of primary aggregates ( $n = 5$  biologically independent samples), which express markers of oRG cells (HOPX and SOX2), IPCs (TBR2) and neurons (CTIP2). Aggregates also had increased cellular stress indicated by *PGK1*, *ARCNI* and *GORASP2* staining. Violin plots show expression level and data distribution for each marker in primary cells, primary cells after organoid transplantation and primary cells after being aggregated together. The expression of *PGK1* and *GORASP2* are increased in post-transplanted primary cells from the organoid as well as in primary cell aggregates. Cell types and physical distribution in the primary aggregate are shown. Scale bar, 50  $\mu$ m. Representative image shown ( $n = 3$  replicates).



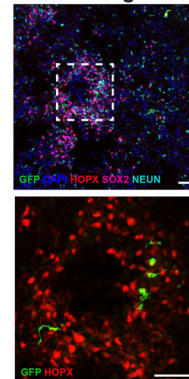
## A FACS Sorting of GFP Positive Cells From Organoids



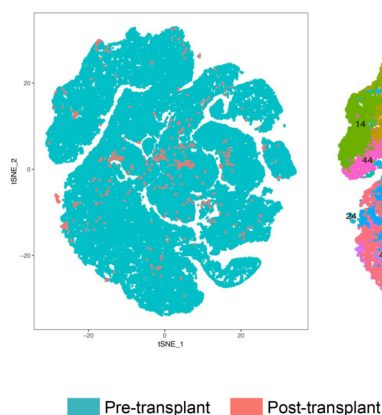
## B Infection Validation GW18 CMV::GFP Adenovirus



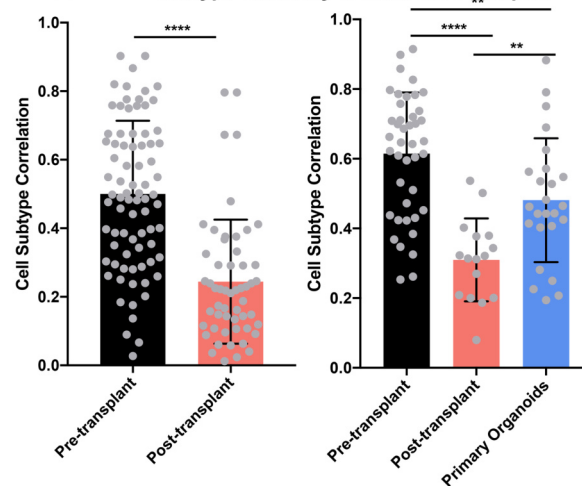
## C GW14 Primary Radial Glia In Week 10 Organoid



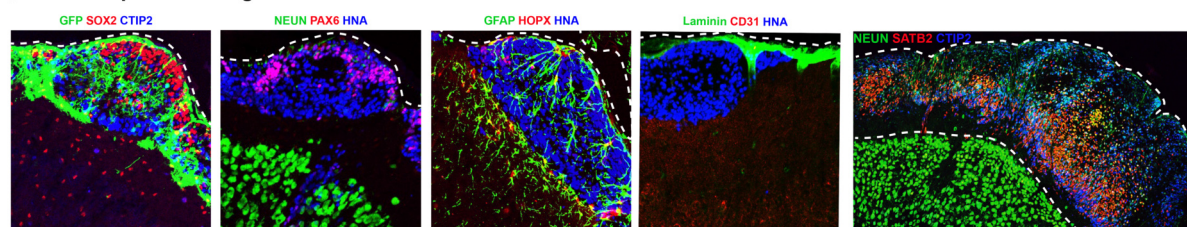
## D Pre- and Post-transplant scRNA-seq



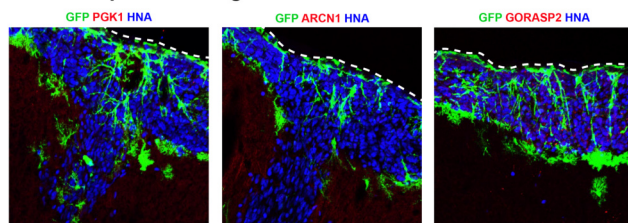
## E Subtype Similarity to Reference Samples



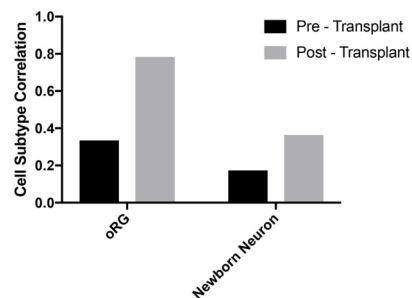
## F Transplanted Organoid Cells Visualized in Mouse Cortex



## G Transplanted Organoid Cells have Reduced Cellular Stress



## H Organoid Cell Subtypes After Mouse Transplant



Extended Data Fig. 14 | See next page for caption.

**Extended Data Fig. 14 | Organoid transplantation at multiple time points.**

**a**, FACS plots showing dummy infection (left) and transplanted organoids (right) in terms of their GFP signal (x-axis) versus sidescatter (y-axis). Cells in the gated region were collected (% of parent written on plot) and sequenced for transplantation 2.5 weeks after incubating in the organoid, representative plot shown on right,  $n = 5$ . **b**, Immunohistochemical validation that cells infected with GFP virus were all SOX2-labelled progenitor in cells dissociated from primary cortical tissue GW14–20. Scale bar, 50  $\mu$ m, representative image shown ( $n = 5$  replicates). **c**, An additional example of primary cell integration into organoids after transplant, in which the primary cells integrate into organoid rosettes ( $n = 7$  primary samples into 21 organoids across 2 independent studies). **d**,  $t$ -SNE of pre- and post-transplant primary cells, as well as the cluster designations. Many cell types represented in pre-transplanted cells are not present in the post-transplant population. **e**, Subtype similarity correlation between pre-transplant, post-transplant, and primary aggregate samples. Includes plot (bar is average subtype correlation, error bars are s.e.) as a replicate of the experiment in Fig. 5b, validating that at older organoid ages (week 12) the post-transplanted cells are still significantly impaired in their subtype specification (\*\*\*\* $P = 1.46 \times 10^{-11}$ ,  $n = 2$  primary biologically

independent samples into 2 organoids in addition to  $n = 5$  biologically independent samples into 10 organoids in Fig. 5, two-sided Welch's  $t$ -test). Primary aggregates are significantly impaired in their subtype specification (\*\* $P = 0.0016$ ), but are significantly better than post-transplanted primary cells (\*\* $P = 0.0037$ ). This may be related to non-neural populations in the aggregates. **f**, Transplanted organoid cells were visualized in the mouse cortex after 2 and 5 weeks post-transplant ( $n = 13$  independent mice transplanted with 14 organoids derived from 2 induced PSC lines across 2 independent experiments). Human cells were visualized by GFP and human nuclear antigen (HNA) expression. Organoid-derived cells expressed markers of progenitors (SOX2 and PAX6), neurons (CTIP2, SATB2 and NEUN) and astrocytes (GFAP and HOPX). Mouse-derived vascular cells (laminin and CD31) innervate the organoid transplant. **g**, After 2 weeks post-transplantation, organoid cells showed reduced expression of the glycolysis gene *PGK1* and ER stress genes *ARCNI* and *GORASP2* ( $n = 6$  transplanted mice stained with each marker independently from 2 induced PSC lines across 2 independent experiments). **h**, Subtype correlation analysis of pre- and post-transplanted organoid cells shows an increase in oRG subtype identity (similarity to primary cluster 26) and in newborn neurons (similarity to primary cluster 22).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used for data collection
Data analysis	Open source software, including cellranger v2, Seurat v2 and simple R correlations (R-3.4.1), and FlowJo v10.6.1 were used in this study for data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Single-cell RNA sequencing data has been deposited in dbGAP for accession "A Cellular Resolution Census of the Developing Human Brain" and in GSE132672. An interactive browser of single cell data and raw and processed counts matrices can be found at the UCSC cell browser website: <https://organoidreportcard.cells.ucsc.edu>. Source data for Figures 1-5 and Extended Data Figures 1-14 are available with the paper. Remaining source data can be retrieved directly from the single-cell data available in public repositories or from the UCSC cell browser website.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by number of biological replicates of distinct individuals and were chosen to be a minimum of three, but potentially more were used. No statistical methods were used to determine sample size.
Data exclusions	No data were excluded
Replication	Reproducibility was ensured by repeating experiments multiple times, including doing experiments on different days, even if distinct individuals were being assayed. All replication attempts were successful.
Randomization	Randomization was performed computationally for subsets of analysis, including WGCNA and variancePartition.
Blinding	No investigators were blinded in this study and blinding was not relevant because we used quantitative methods to measure our results, especially with regards to single-cell analysis

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	Sox2 (Santa Cruz, 1:500, sc-365823), Hopx (Santa Cruz, 1:250, sc-398703), Satb2 (Abcam, 1:250, ab51502), AUTS2 (abcam, 1:100, ab243036), Human Nuclei (Millipore, 1:500, MAB1281), Rabbit: Hopx (Proteintech, 1:500, 11419-1-AP), GORASP2 (Proteintech, 1:50, 10598-1-AP), ARCN1 (Proteintech, 1:50, 23843-1-AP), PGK1 (Thermo Fisher 1:50, PA5-13863), PTPRZ1 (Atlas, 1:250, HPA015103), NR2F1 (Novus, 1:100, NBP1-31259), Rat: Ctbp2 (Abcam, 1:500, ab18465), Sheep: Eomes (R&D, 1:200, AF6166), Guinea pig: NeuN (Millipore, 1:500, ABN90), Chicken: Gfp (Aves, 1:500, GFP-1020).
Validation	Only previously published antibodies, or antibodies with company based validation of correct size were used. Website of each company can be consulted.

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	hESC lines: H1 - obtained from WiCell. iPSC lines: 1323_4, WTC10 from the Gladstone Institute, H28126 from University of Chicago
Authentication	Lines were obtained using MTA approval; at time of generation lines were karyotyped for normal identity.
Mycoplasma contamination	Lines are negative for mycoplasma.
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No ICLAC lines were used

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Postnatal day four (p4) male and female NSG (NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ, stock No:005557) mice were used for this study.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	Mouse experiments were approved by UCSF Institutional Animal Care and Use Committee (IAUCUC) protocol AN178775-01.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Human tissue samples were collected without any identifying information including sex or race.
Recruitment	No human participants were used in this study; human tissue was collected from elective terminations with the patients prior consent.
Ethics oversight	All primary tissue was obtained and processed as approved by the UCSF Human Gamete, Embryo and Stem Cell Research Committee (GESCR) approval 10-05113

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation	Transplanted organoids were dissociated as described in the Methods.
Instrument	Becton Dickinson FACSAria
Software	FlowJo
Cell population abundance	1-2% of cells were GFP positive
Gating strategy	Negative control cells were used to set gates for 488, identifying cells above that threshold only

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.



# Targeting of temperate phages drives loss of type I CRISPR–Cas systems

<https://doi.org/10.1038/s41586-020-1936-2>

Received: 1 February 2019

Accepted: 25 November 2019

Published online: 22 January 2020

Clare Rollie<sup>1,6\*</sup>, Anne Chevallereau<sup>1,6\*</sup>, Bridget N. J. Watson<sup>1</sup>, Te-yuan Chyou<sup>2</sup>, Olivier Fradet<sup>1</sup>, Isobel McLeod<sup>1</sup>, Peter C. Fineran<sup>3,4</sup>, Chris M. Brown<sup>2,4</sup>, Sylvain Gandon<sup>5</sup> & Edze R. Westra<sup>1\*</sup>

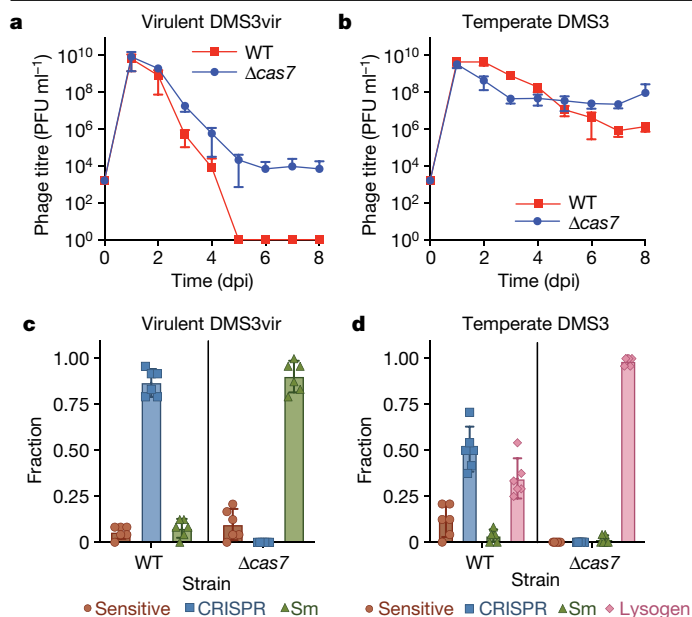
On infection of their host, temperate viruses that infect bacteria (bacteriophages; hereafter referred to as phages) enter either a lytic or a lysogenic cycle. The former results in lysis of bacterial cells and phage release (resulting in horizontal transmission), whereas lysogeny is characterized by the integration of the phage into the host genome, and dormancy (resulting in vertical transmission)<sup>1</sup>. Previous co-culture experiments using bacteria and mutants of temperate phages that are locked in the lytic cycle have shown that CRISPR–Cas systems can efficiently eliminate the invading phages<sup>2,3</sup>. Here we show that, when challenged with wild-type temperate phages (which can become lysogenic), type I CRISPR–Cas immune systems cannot eliminate the phages from the bacterial population. Furthermore, our data suggest that, in this context, CRISPR–Cas immune systems are maladaptive to the host, owing to the severe immunopathological effects that are brought about by imperfect matching of spacers to the integrated phage sequences (prophages). These fitness costs drive the loss of CRISPR–Cas from bacterial populations, unless the phage carries anti-CRISPR (acr) genes that suppress the immune system of the host. Using bioinformatics, we show that this imperfect targeting is likely to occur frequently in nature. These findings help to explain the patchy distribution of CRISPR–Cas immune systems within and between bacterial species, and highlight the strong selective benefits of phage-encoded acr genes for both the phage and the host under these circumstances.

CRISPR–Cas adaptive immune systems provide sequence-specific resistance against phage infections by inserting phage-derived sequences (spacers) of around 30 bp into CRISPR loci on the host genome<sup>4</sup>. Upon reinfection, CRISPR transcripts guide Cas proteins to destroy the matching target<sup>5</sup>. The uptake of new spacers is far more efficient if a pre-existing spacer has partial complementarity to the phage. This process (known as ‘priming’) is widespread in type I CRISPR–Cas systems, and provides protection against phage mutants that overcome host resistance by point mutation of their target sites<sup>6,7</sup>. However, partially matching spacers can also cause immunopathological effects when temperate phages enter the lysogenic lifecycle, during which the phage genome is integrated into that of the host and exists in a dormant state until it is induced. Lysogens (that is, bacterial cells that carry at least one prophage) are found across bacterial genera; however, their frequencies vary<sup>8</sup>. During lysogeny, a primed CRISPR–Cas immune system may target the partially complementary site in the prophage, causing damage to both the phage and host DNA and resulting in the induction of the SOS response<sup>9</sup>. However, whether and how these potentially negative effects influence the evolutionary and population dynamics of CRISPR–phage interactions remains unclear, because these processes have previously been studied only in the context of virulent phages<sup>2,3</sup>.

## Temperate phages persist despite CRISPR

To explore CRISPR–phage interactions in the context of lysogeny, we infected *Pseudomonas aeruginosa* PA14 with the temperate phage DMS3 or the virulent mutant DMS3vir (a DMS3 mutant that is locked in the lytic cycle through mutation of the c-repressor gene) and monitored bacterial and phage population dynamics for eight days. *P. aeruginosa* strain PA14 carries a type I-F CRISPR–Cas immune system, in which spacer 1 of CRISPR array 2 has an imperfect match (5 mismatches) to gene 42 of DMS3<sup>9,10</sup>. As previously reported<sup>3</sup>, DMS3vir was driven extinct by wild-type bacteria at five days post-infection owing to the evolution of CRISPR-based resistance, whereas a  $\Delta cas7$  mutant of the PA14 strain (hereafter  $\Delta cas7$ —which lacks a functional CRISPR–Cas system—evolved surface-based resistance that allowed for phage persistence (Fig. 1a–c). By contrast, both wild-type and  $\Delta cas7$  bacteria were unable to clear wild-type DMS3 infections (Fig. 1b), which suggests that the ability to transmit vertically (during lysogeny) is a critical determinant of the survival of temperate phages when bacteria encode CRISPR–Cas immune systems. To understand the evolutionary drivers of these population dynamics, we isolated bacterial clones from the DMS3-infected cultures at three days post-infection and quantified the proportion of lysogens and bacteria with CRISPR-based and surface-based

<sup>1</sup>ESI, Biosciences, University of Exeter, Penryn, UK. <sup>2</sup>Department of Biochemistry, University of Otago, Dunedin, New Zealand. <sup>3</sup>Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand. <sup>4</sup>Genetics Otago, University of Otago, Dunedin, New Zealand. <sup>5</sup>CEFE, Université de Montpellier, CNRS, EPHE, IRD, Université Paul Valéry Montpellier 3, Montpellier, France. <sup>6</sup>These authors contributed equally: Clare Rollie, Anne Chevallereau. \*e-mail: C.Rollie@exeter.ac.uk; A.Chevallereau@exeter.ac.uk; E.R.Westra@exeter.ac.uk

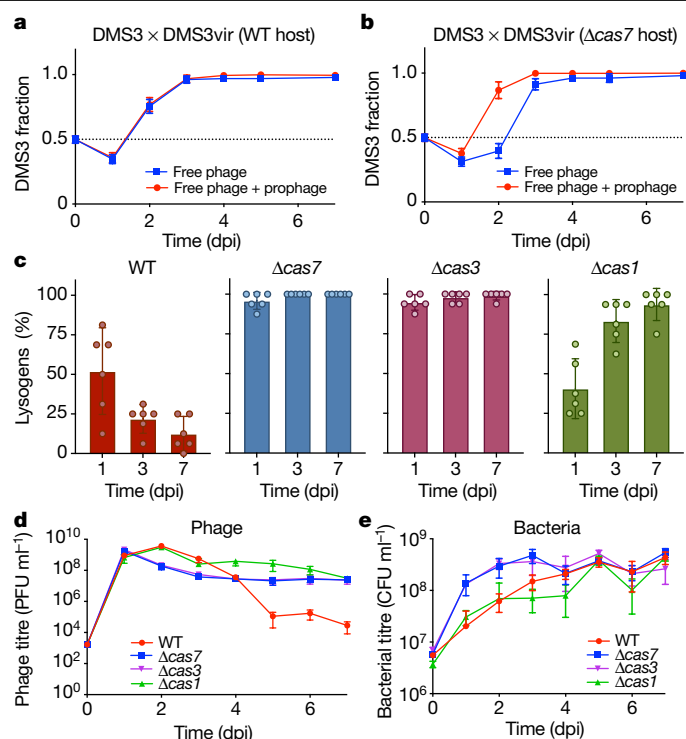


**Fig. 1 | Phage persistence and evolution of resistance in the host upon infection with virulent or temperate phages.** **a, b**, Phage densities over time after infection of wild-type (WT) PA14 or the  $\Delta cas7$  mutant with DMS3vir (a) or DMS3 (b). The limit of phage detection is 200 plaque-forming units per millilitre (PFU ml<sup>-1</sup>). **c, d**, Fraction of bacteria that evolved resistance at three days post-infection (dpi) after infection with the phages DMS3vir (c) or DMS3 (d), either through CRISPR–Cas (CRISPR), surface modification (sm) or lysogeny (lysogen). Fractions are based on 24 random clones per replicate experiment. In all panels, data are the mean of six biologically independent replicates per treatment. Error bars represent 95% confidence intervals.

resistance. This showed that the evolution of CRISPR-based resistance was substantially reduced in wild-type populations of bacteria exposed to DMS3 compared to those exposed to DMS3vir (Fig. 1c, d); instead, many wild-type bacteria carried the DMS3 prophage (Fig. 1d), which confers phage resistance through superinfection exclusion. However, lysogeny levels were reduced in wild-type hosts compared to in the  $\Delta cas7$  strain, which almost invariably carried the prophage (Fig. 1d).

The difference in DMS3 and DMS3vir persistence could be due to differences in the evolution of resistance in the host or to the ability of DMS3 to transmit vertically while continuously releasing free phage particles through prophage induction. To distinguish between these possibilities, we infected wild-type and  $\Delta cas7$  strains with an equal mix of DMS3 and DMS3vir, and observed bacterial and phage population dynamics and resistance evolution similar to those that were observed during infection with the temperate phage alone (Extended Data Fig. 1). Next, we examined how the relative frequencies of DMS3 and DMS3vir changed over time (Fig. 2a, b). Initially, DMS3vir outcompetes DMS3, consistent with the idea that high densities of sensitive hosts favour horizontal transmission (the lytic replication cycle)<sup>11</sup>. However, at later time points, all free phage particles belong to the DMS3 phage genotype. As the host population that was experienced by the two phages was identical, this demonstrates that vertical transmission facilitates the observed persistence of DMS3. The persistence of free phages was facilitated despite the low frequencies of lysogeny in wild-type compared to  $\Delta cas7$  bacteria (compare the blue and red lines in Fig. 2a, b, Extended Data Fig. 1c, d).

To understand why lysogen formation was depressed in wild-type bacteria compared to  $\Delta cas7$  bacteria, we performed temporal sampling of DMS3-infected populations over seven days. This revealed that lysogeny in wild-type bacteria was already depressed at one day post-infection and continued to decline until seven days post-infection, whereas the proportion of lysogens in isogenic mutants with



**Fig. 2 | The effect of CRISPR adaptation and interference on lysogeny and phage persistence.** **a, b**, Relative frequencies of DMS3 over time after infection of wild-type PA14 (a) or the  $\Delta cas7$  mutant (b) host with an equal mix of DMS3 and DMS3vir. Relative frequencies are shown both for the free and total (that is, including lysogens) phage population. **c**, Percentage of DMS3 lysogens in the host population at 1, 3 or 7 days post-infection of the wild-type PA14 strain, the isogenic CRISPR-interference deficient mutants  $\Delta cas7$  and  $\Delta cas3$ , or the isogenic CRISPR-adaptation-deficient mutant  $\Delta cas1$ , based on 24 random clones per replicate experiment per time point. **d, e**, DMS3 phage (d) and bacterial densities (e) during this co-culture experiment. CFU, colony-forming unit. In all panels, data are the mean of six biologically independent replicates per treatment. Error bars represent 95% confidence intervals.

a defective CRISPR-interference pathway (the  $\Delta cas3$  and  $\Delta cas7$  mutants of PA14) was high and constant during this same period (Fig. 2c). Crucially, the proportion of DMS3 lysogens at one day post-infection was also depressed in a  $\Delta cas1$  mutant of PA14, which is unable to acquire novel spacers (adaptation) but is proficient in detecting and destroying complementary DNA (interference). However, in this background, the proportion of lysogens increased between one and seven days post-infection to levels similar to those in the CRISPR-interference mutants (Fig. 2c). These data therefore suggest that wild-type and  $\Delta cas1$  bacteria are partially resistant to DMS3 infection (even in the absence of spacer acquisition), which results initially in fewer lysogens and further reductions if the hosts carry the genetic machinery to acquire additional spacers.

### Mismatched spacers are maladaptive

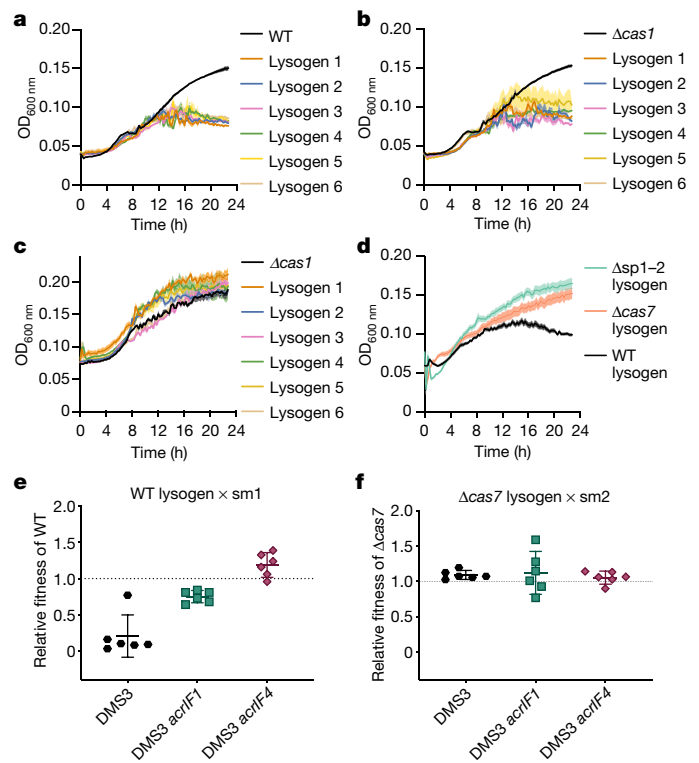
We hypothesized that the observed suppression of lysogeny in wild-type and  $\Delta cas1$  backgrounds was dependent on the imperfect match (5 mismatches) between gene 42 of DMS3 and spacer 1 of CRISPR array 2 (refs. <sup>9,10</sup>). To test this, we infected a strain that lacks CRISPR array 2 ( $\Delta CRISPR2$ ) with DMS3 and observed that the plasmid-based expression of spacer 1 ( $\Delta CRISPR2$ -sp1) led to population dynamics similar to those observed during infection of the wild-type strain (Extended Data Fig. 2a, b) and suppression of lysogeny (Extended Data Fig. 2c), whereas expression of a control non-targeting spacer ( $\Delta CRISPR2$ -NT) did not.

To further corroborate the hypothesis that the suppression of lysogeny was caused by partial CRISPR–Cas resistance, we performed infections with phages that carry *acr* genes, which are widespread and diverse genes (currently classified into 47 families) that are encoded by a range of mobile genetic elements and that block CRISPR interference<sup>12</sup>. As expected, the infection of wild-type bacteria with a mutant DMS3 phage that carries *acrIF1* or *acrIF4* (both of which block the type I-F CRISPR–Cas system of PA14<sup>13</sup>) showed levels of lysogeny similar to those observed for interference-deficient mutants (Extended Data Fig. 3a).

Although lysogen formation was reduced in wild-type and *ΔcasI* bacteria, the concentration of free phage was actually higher during the early stages of DMS3 infection in these hosts (Fig. 2d). This effect disappeared by seven days post-infection if bacteria were unable to acquire new spacers (the *ΔcasI* strain), or inverted at four days post-infection if the bacteria had a functional CRISPR–Cas system (Fig. 2d). This also coincided with reduced bacterial densities of wild-type and *ΔcasI* populations compared to CRISPR-interference mutants during the early stages of the phage epidemic (Fig. 2e), which suggests that functional CRISPR–Cas immune systems may be maladaptive in this context. We speculated that the high DMS3 titres during early stages of infection of wild-type and *ΔcasI* bacteria (Figs. 1b, 2d), and the concurrent reduced bacterial densities, could be due to interactions between the CRISPR-interference machinery and the partially complementary prophage (autoimmunity), which can activate an SOS response<sup>9</sup> and which—in turn—may trigger prophage induction<sup>14</sup>. To test this, we isolated six independent DMS3 lysogens from wild-type, *Δcas7* and *ΔcasI* backgrounds at one day post-infection. Growth measurements revealed considerable fitness costs of lysogeny in wild-type and *ΔcasI* backgrounds, whereas the growth of *Δcas7* lysogens was unaffected (Fig. 3a–c). The growth of lysogens that lack spacers 1 and 2 of CRISPR 2 (*Δsp1–2*) or lack the entire CRISPR 2 array (*ΔCRISPR2-NT*) was comparable to that of *Δcas7* lysogens (Fig. 3d, Extended Data Fig. 2c), but restoring the expression of CRISPR 2 spacer 1 (*ΔCRISPR2-sp1*) resulted in a reduced growth rate (Extended Data Fig. 2d), which confirms that this spacer is responsible for the reduced fitness of wild-type bacteria that carry the prophage. Competition of lysogens against bacteria with surface-based resistance confirmed a fitness cost of lysogeny in the wild-type—but not in the *Δcas7*—background (relative fitness < 1, one-tailed Wilcoxon signed-rank and *t*-tests; wild type, degrees of freedom = 5,  $P = 0.016$ ; *Δcas7*,  $t_5 = 3.6551$ ,  $P = 0.99$ ) (Fig. 3e, f). This effect was reduced if the prophage encoded an *Acr* (one-tailed *t*-test; *acrIF1*,  $t_5 = 16.562$ ; *acrIF4*,  $t_5 = 14.805$  and  $P < 0.0001$  in both cases) (Fig. 3e), whereas *acr* genes had no effect on fitness of *Δcas7* lysogens (analysis of variance (ANOVA)  $F_{2,15} = 0.205$ ,  $P = 0.82$ ) (Fig. 3f).

Next, we estimated prophage induction levels in wild-type and *Δcas7* lysogens. To this end, we pelleted bacterial cultures and washed away free phages, and then resuspended in fresh medium, and monitored bacterial densities and free phage accumulation over 25 h. Although initial cell densities were similar for all strains, the growth of wild-type cultures plateaued at a lower density than that of other backgrounds (Extended Data Fig. 3b), but wild-type cultures accumulated more phage particles compared to *Δcas7* lysogens (Extended Data Fig. 3c)—unless the prophage carried an *acr* gene. Taken together, these data support the hypothesis that (unless DMS3 encodes *acr* genes) CRISPR–Cas causes immunopathology during vertical transmission of the phage, which triggers prophage induction and therefore explains the high free-phage titres during the early infection stage.

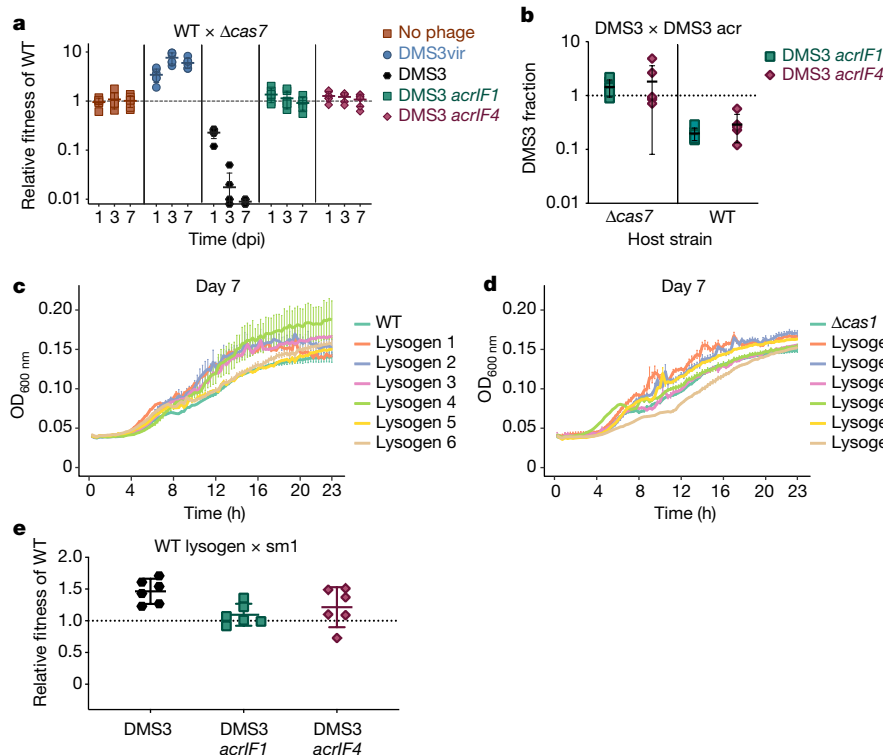
These immunopathological effects may select for CRISPR mechanisms that selectively target phages during their lytic cycle, as has previously been described for type III systems<sup>15,16</sup>. However, competition between DMS3 and DMS3vir phages showed that the main determinant of their relative fitness is the proportion of sensitive bacteria in the population<sup>11</sup> (two-way ANOVA,  $F_{6,83} = 3.1683$ ,  $P = 0.008$ ) (Extended Data Fig. 3d) and is independent of whether resistance of bacteria in the population was CRISPR-based or surface-based ( $F_{1,76} = 2.8923$ ,  $P = 0.09$ )



**Fig. 3 | Fitness of lysogens with an active CRISPR–Cas system is reduced unless they encode *acr* genes.** **a**, Twenty-four-hour growth curves of uninfected control cultures, or six independent DMS3 lysogens in wild-type PA14 (**a**), *ΔcasI* (CRISPR-adaptation-deficient) (**b**) and *Δcas7* (CRISPR-interference-deficient) (**c**) genetic backgrounds. Lysogens were isolated from day 1 of the co-culture experiment shown in Fig. 2. Curves are the mean of six (**a**, **b**) or four (**c**) replicates and shaded areas represent s.e.m. OD<sub>600 nm</sub>, optical density at 600 nm. **d**, Twenty-four-hour mean growth curves of six lysogens in wild-type, *Δcas7* and *Δsp1–2* (carrying a deletion of CRISPR2 spacers 1 and 2) backgrounds isolated from six biological replicates. Each growth curve was performed in five technical replicates. Shaded areas represent s.e.m. **e**, **f**, Fitness relative to a mutant of PA14 with a surface modification (sm) of wild-type PA14 lysogens isolated one day post-infection with DMS3 (**e**) or PA14 *Δcas7* lysogens isolated one day post-infection with DMS3 (**f**). Relative fitness was determined after one day of competition. Each point represents the average relative fitness of one independent lysogen clone measured across six biologically independent experiments. Error bars indicate 95% confidence intervals.

(Extended Data Fig. 3d). This suggests that the type I CRISPR–Cas system of *P. aeruginosa* lacks the ability to distinguish between phages that enter lytic or lysogenic cycles.

Given these opposing fitness effects of CRISPR–Cas immune systems during horizontal<sup>2,3</sup> and vertical transmission of DMS3 (Fig. 3), it is unclear what the net fitness effects of CRISPR–Cas systems are during temperate-phage infections. To explore this, we competed wild-type and *Δcas7* strains in the presence of DMS3vir, DMS3, DMS3 *acrIF1* or DMS3 *acrIF4* phages and determined their relative fitness at days 1, 3 and 7. Notably, although wild-type bacteria were fitter than *Δcas7* in the presence of the virulent phage (one-tailed Wilcoxon signed-rank test, relative fitness > 1,  $P = 0.016$ ) (Fig. 4a), they were considerably less fit during temperate-phage infection (relative fitness < 1,  $P = 0.018$ ) (Fig. 4a), unless the temperate phage carried *acr* genes (relative fitness ≠ 1,  $P = 0.16$ ) (Fig. 4a). Therefore, in this context, *acr* genes not only provide a benefit to the phage during both horizontal<sup>3,17,18</sup> and vertical transmission of the phage (one-tailed *t*-test, relative fitness ≠ 1; *acrIF1*,  $t_5 = 39.30$ ,  $P < 0.0001$ ; *acrIF4*,  $t_5 = 11.61$ ,  $P < 0.0001$ ) (Fig. 4b) but also to the host, by preventing autoimmunity.



**Fig. 4 | Lysogens evolve to mitigate fitness costs.** **a**, Relative fitness of wild-type PA14 during competition with the  $\Delta cas7$  mutant strain following infection with  $10^4$  PFU of DMS3, the lytic mutant DMS3vir or the anti-CRISPR-encoding mutants DMS3 *acrIF1* and DMS3 *acrIF4*. **b**, Relative fitness of DMS3 (free phages + lysogens) following three days of competition with DMS3 *acrIF1* or DMS3 *acrIF4* on either wild-type PA14 or  $\Delta cas7$  mutant. **c**, **d**, Twenty-four-hour growth curves of uninfected control cultures, or six biologically independent DMS3 lysogens in the wild-type PA14 (c) or  $\Delta cas1$  (d) genetic backgrounds,

which were isolated from day 7 of the co-culture experiment shown in Fig. 2. Curves are the mean of six replicates and error bars represent s.e.m. **e**, Relative fitness of a DMS3 lysogen in a wild-type PA14 genetic background, isolated from day 5 of a co-culture experiment, during competition with a surface mutant. Relative fitness was calculated after one day of competition. All panels show the mean of six biologically independent replicates per treatment. Error bars represent 95% confidence intervals (a, b, e) or s.e.m. (c, d).

## Bacteria evolve to lose CRISPR–Cas

We next monitored whether wild-type lysogens evolved to alleviate these autoimmunity costs through mutation of their immune system or the prophage. Lysogens in wild-type and adaptation-deficient  $\Delta cas1$  backgrounds, which had reduced growth during early infection, were selected from a late stage of the infection (seven days after infection). Growth curves revealed that the negative fitness consequences of CRISPR–Cas immune systems had disappeared in these ‘late’ lysogens, and that these bacteria now had growth rates comparable those of ancestral hosts that lack the prophage (Fig. 4c, d; compare to Fig. 3a, b). A competition experiment confirmed this finding, which shows that late lysogens in a wild-type background were fitter than a surface mutant (one-tailed *t*-test, relative fitness  $> 1$ ,  $t_5 = 5.985$ ,  $P < 0.001$ ) (Fig. 4e), and the presence of *acr* genes in the prophage no longer increased the fitness of the host (one-tailed *t*-test, *acrIF1*,  $t_5 = -5.5085$ ,  $P = 1.00$ ; *acrIF4*,  $t_5 = -2.0395$ ,  $P = 0.95$ ) (Fig. 4e, compare to Fig. 3e).

To understand the mechanistic basis for this alleviation in fitness costs, we performed PCR analyses of the CRISPR loci of six independent lysogens in wild-type,  $\Delta cas1$  and  $\Delta cas7$  backgrounds, isolated from early or late time points after infection. Both CRISPR loci (1 and 2) amplified as expected in early lysogens (Extended Data Fig. 4a); however, amplification failed in many late lysogens in wild-type and  $\Delta cas1$  backgrounds (Extended Data Fig. 4a, negative PCR indicated by red frame), which suggests that the CRISPR–Cas locus was lost in these clones. Whole-genome sequencing of late lysogens revealed large genomic deletions (between about 50 and 230 kb), associated with prophage integration, that encompassed the entire CRISPR–Cas locus—but contained

no essential gene<sup>19</sup>—in wild-type and  $\Delta cas1$  backgrounds, whereas the genome remained intact in  $\Delta cas7$  late lysogens (Extended Data Fig. 4c–e, Extended Data Table 1). This loss of the CRISPR–Cas locus was clearly driven by the immunopathological effects of CRISPR 2 spacer 1, as the locus was maintained in lysogens that lack CRISPR 2 ( $\Delta CRISPR2$ ) but was lost when expression of spacer 1 was restored ( $\Delta CRISPR2$ -sp1) (Extended Data Fig. 2b). The loss of CRISPR–Cas was also avoided when wild-type bacteria were lysogenized by DMS3 that carry *acr* genes (Extended Data Fig. 4a).

## Generality of the empirical data

Collectively, these data show that temperate-phage infection can drive the rapid loss of CRISPR–Cas immune systems from bacterial genomes owing to immunopathological effects that manifest during vertical transmission. To generalize our findings beyond phage DMS3 (which belongs to the *Caudovirales*), we introduced a priming spacer with one mismatch against the PF5 prophage (which belongs to the *Inoviridae*) that is naturally present in the genome of wild-type PA14<sup>20</sup>. Expression of a PF5-priming spacer caused reduced growth of the wild-type PA14 strain and strong selection for bacteria that carry deletions in their CRISPR–Cas immune system (Extended Data Fig. 5). These observations therefore suggest that immunopathological effects generally drive selection for CRISPR loss, whenever hosts carry spacers primed against their prophage. Indeed, when we formalized these ideas in a theoretical framework, we recovered population and evolutionary dynamics very similar to those observed in our experiments (Extended Data Fig. 6; a detailed description of the model is provided in the Supplementary Information).



Crucially, in nature, priming is thought to be frequent owing to the relaxed sequence-identity requirements for triggering this pathway: up to 13 mismatches between a pre-existing spacer and the target are tolerated<sup>7</sup>. Indeed, analysis of the spacers from more than 170,000 bacterial genomes and a dataset of about 20,000 prophages<sup>21</sup> revealed that pre-existing spacers with perfect (no mismatches) or imperfect (1–5 mismatches) targets to temperate phages are common within genera (Extended Data Fig. 7; further information is in the Supplementary Information). On average, 49% of all prophages were targeted by priming spacers carried by bacteria within the same genera as the lysogen (Extended Data Fig. 7c). If mismatched spacers generally cause residual DNA cleavage activity by the CRISPR immune system, they would be expected to cause immunopathological effects similar to those observed for the *P. aeruginosa*–DMS3 interaction. Indeed, when all complete *P. aeruginosa* genomes were assessed, we observed a significant enrichment for an increase in the frequency of *acr* genes when self-targeting spacers matched prophages (Extended Data Fig. 8), which is consistent with the idea that these spacers cause immunopathological effects. Together, these bioinformatic analyses suggest that—in nature—the maladaptive effects of CRISPR–Cas against temperate phages are likely to be common.

In natural ecosystems, bacteria are frequently exposed to both temperate and virulent phages, and the latter may superinfect the lysogens. To further generalize our findings, we explored whether these superinfections would be likely to change the observed population and evolutionary dynamics. We found that a mixed infection of wild-type PA14 with the temperate phage DMS3 and the virulent phage LMA2 (which can superinfect DMS3 lysogens) resulted in rates of lysogenization similar to those with the temperate phage alone (Extended Data Fig. 9d, e) and high fitness costs of CRISPR–Cas (Extended Data Fig. 9g), which drive the evolutionary loss of the immune systems (Extended Data Fig. 9f). Our theoretical model recovered similar dynamics, independent of whether CRISPR-based resistance evolves against only one (as in our experiment) or against both phages (Extended Data Fig. 9h–o).

## Discussion

The observations that approximately 60% of all bacterial genotypes lack CRISPR–Cas systems, and that closely related strains differ in whether they encode *cas* genes, suggest that these systems are frequently gained and lost from bacterial genomes<sup>22</sup>. Although phage infection is typically assumed to be an important selective force for the maintenance of CRISPR–Cas immune systems, this work reveals their rapid loss when exposed to a temperate phage, as a result of the immunopathological effects that occur when the CRISPR immune system is primed against this phage. Some CRISPR–Cas immune systems may mitigate some of these autoimmunity costs by limiting the spacer acquisition from, and cleavage of, transcriptionally silent elements<sup>15,23</sup>. However, CRISPR-based autoimmunity has frequently been reported<sup>24–27</sup>, and the high frequencies at which primed self-targeting interactions occur suggests that this is probably a major driver of the evolutionary loss of CRISPR systems in nature.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1936-2>.

- Stewart, F. M. & Levin, B. R. The population biology of bacterial viruses: why be temperate. *Theor. Popul. Biol.* **26**, 93–117 (1984).
- Westra, E. R. et al. Parasite exposure drives selective evolution of constitutive versus inducible defense. *Curr. Biol.* **25**, 1043–1049 (2015).
- van Houte, S. et al. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385–388 (2016).
- Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
- Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
- Datsenko, K. A. et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).
- Fineran, P. C. et al. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl Acad. Sci. USA* **111**, E1629–E1638 (2014).
- Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J.* **11**, 1511–1520 (2017).
- Heussler, G. E. et al. Clustered regularly interspaced short palindromic repeat-dependent, biofilm-specific death of *Pseudomonas aeruginosa* mediated by increased expression of phage-related genes. *mBio* **6**, e00129-15 (2015).
- Zegans, M. E. et al. Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *J. Bacteriol.* **191**, 210–219 (2009).
- Berngruber, T. W., Froissart, R., Choisy, M. & Gandon, S. Evolution of virulence in emerging epidemics. *PLoS Pathog.* **9**, e1003209 (2013).
- Trasnidou, D. et al. Keeping CRISPR in check: diverse mechanisms of phage-encoded anti-CRISPRs. *FEMS Microbiol. Lett.* **366**, fnz098 (2019).
- Bondy-Denomy, J. et al. Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature* **526**, 136–139 (2015).
- Little, J. W. & Michalowski, C. B. Stability and instability in the lysogenic state of phage lambda. *J. Bacteriol.* **192**, 6064–6076 (2010).
- Goldberg, G. W., Jiang, W., Bikard, D. & Maraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR–Cas targeting. *Nature* **514**, 633–637 (2014).
- Samai, P. et al. Co-transcriptional DNA and RNA cleavage during type III CRISPR–Cas immunity. *Cell* **161**, 1164–1174 (2015).
- Landsberger, M. et al. Anti-CRISPR phages cooperate to overcome CRISPR–Cas immunity. *Cell* **174**, 908–916 (2018).
- Borges, A. L. et al. Bacteriophage cooperation suppresses CRISPR–Cas3 and Cas9 immunity. *Cell* **174**, 917–925 (2018).
- Poulsen, B. E. et al. Defining the core essential genome of *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **116**, 10072–10080 (2019).
- Mooij, M. J. et al. Characterization of the integrated filamentous phage Pf5 and its involvement in small-colony formation. *Microbiology* **153**, 1790–1798 (2007).
- Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Res.* **39**, W347–W352 (2011).
- Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
- Levy, A. et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).
- Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* **26**, 335–340 (2010).
- Vercoe, R. B. et al. Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet.* **9**, e1003454 (2013).
- Jiang, W. et al. Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet.* **9**, e1003844 (2013).
- Goldberg, G. W. et al. Incomplete prophage tolerance by type III-A CRISPR–Cas systems reduces the fitness of lysogenic hosts. *Nat. Commun.* **9**, 61 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020



# Article

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Bacterial strains and viruses

*P. aeruginosa* UCBPP-PA14 (referred to as wild type) and *P. aeruginosa* UCBPP-PA14 *csy3::lacZ* (referred to as  $\Delta cas7$ ) were used in all experiments, and have previously been described<sup>28</sup>. The surface mutant *Tn::pilA* (which lacks a pilus, and is referred to as sm2) has previously been described<sup>29</sup> and was used in competition experiments with a bacteriophage insensitive mutant (BIM) with two additional acquired spacers over those present in the wild-type strain, as previously described<sup>2</sup>.

Lysogens in wild-type or CRISPR-mutant backgrounds used in growth curves, fitness and induction experiments were generated in this study; the presence of a prophage was confirmed by PCR. PA14 strains  $\Delta cas3$ ,  $\Delta cas1$ ,  $\Delta CRISPR2$  and  $CRISPR2 \Delta sp1-2$  identified, respectively, as SMC4268, SMC4277 and SMC3895 and SMC4707, have previously been described<sup>10,28</sup>.

Phage amplifications were carried out on *P. aeruginosa* UCBPP-PA14 *csy3::lacZ*. DMS3 and the obligate lytic variant DMS3vir have previously been described<sup>30</sup>. DMS3 *acrIF1* and DMS3 *acrIF4* phages were generated in this study following a method for hybrid phage construction detailed in a previous study<sup>17</sup>. DMS3vir *acrIF1* was used in downstream analyses and has previously been described<sup>3</sup>. A virulent phage capable of infecting DMS3 lysogens (LMA2, which has previously been described<sup>31</sup>) was used in co-culture experiments.

All bacterial strains were grown at 37 °C in LB broth or M9 medium (22 mM Na<sub>2</sub>HPO<sub>4</sub>; 22 mM KH<sub>2</sub>PO<sub>4</sub>; 8.6 mM NaCl; 20 mM NH<sub>4</sub>Cl; 1 mM MgSO<sub>4</sub>; and 0.1 mM CaCl<sub>2</sub>) supplemented with 0.2% glucose. When appropriate, medium was further supplemented with gentamycin (50 mg/ml) and arabinose (1% w/v).

### Evolution experiments

To monitor the evolution of bacterial resistance in response to phage infection and the associated bacterial and phage population dynamics, microcosms with 6 ml of M9 medium supplemented with 0.2% glucose were inoculated with approximately 10<sup>6</sup> CFU bacteria from fresh overnight cultures of the corresponding bacterial strains. These cultures were infected with 10<sup>4</sup> PFU of DMS3vir, DMS3, DMS3 *acrIF* or LMA2 phages, followed by incubation at 37 °C and shaking at 180 rpm. Cultures were transferred 1:100 into fresh medium every 24 h for 5–8 days. All experiments were performed in six independent replicates.

### Determination of resistance phenotypes

Resistance phenotypes were determined at day 3 or day 7 by streaking individual colonies (24 randomly picked per replicate) through DMS3vir and anti-CRISPR phage DMS3vir *acrIF1*. Surface modification was confirmed by colony morphology, broad-range resistance to DMS3vir and DMS3vir *acrIF1*, and a lack of newly acquired spacers. Lysogens were determined by broad resistance to both phage and a positive PCR amplification of the c-repressor gene in bacterial genome, amplified using primers 5'-GCCGAATGAGCGCTAAACC-3' and 5'-CAAGTGCTTACGAGGAATGC-3'. CRISPR resistance was confirmed by resistance to DMS3vir, but not to DMS3vir *acrIF1* and a PCR confirming spacers had been added to one of the CRISPR arrays. Primers 5'-CTAAGCCTGTACGAAGTCTC-3' and 5'-CGCGAAGGCCAGCGCGCGGTG-3' were used to amplify CRISPR array 1, and primers 5'-GCCGTCCAGAAGTCAACACCG-3' and 5'-TCAGCAAGTTACGAGACCTCG-3' for CRISPR array 2. As a positive control for PCR, primers 5'-GCTTGACAGTTCCTCAACGAG-3' and 5'-CACCAGGAAATTCAGGTAGGG-3' were used to amplify the housekeeping control gene *fimV*, which encodes a protein that is involved in pilus formation.

### Bacterial and phage titres

Bacterial densities were determined by plating on LB agar dilutions of samples taken at each transfer in M9 salts (22 mM Na<sub>2</sub>HPO<sub>4</sub>; 22 mM KH<sub>2</sub>PO<sub>4</sub>; 8.6 mM NaCl; 20 mM NH<sub>4</sub>Cl; 1 mM MgSO<sub>4</sub>; and 0.1 mM CaCl<sub>2</sub>). Phages were extracted at each transfer by chloroform extraction (sample:chloroform 10:1, v/v), and phage titres were determined by spotting serial dilutions of isolated phage samples in M9 salts on a lawn of PA14  $\Delta cas7$ .

### Phage competition

**Fixed phenotypes.** Competition experiments were performed in glass vials in 6 ml of M9 medium. Experiments were initiated by inoculating a 1:100 dilution of different mixes of overnight cultures of the  $\Delta cas7$  strain and either the surface mutant  $\Delta pilA^{29}$  (sm2) or BIM strain (2 spacers targeting DMS3 phage). A 1:1 mix of the phages DMS3vir and DMS3 was added (10<sup>8</sup> PFU) and the vials were incubated at 37 °C with shaking for 8 h. Phages were extracted by chloroform extraction and spot assays carried out to determine phage titre. To determine phage ratios (DMS3:DMS3vir), plaque assays were carried out by serially diluting phage extractions and adding 200 µl of the selected dilution to 600 µl of  $\Delta cas7$  overnight culture and 6 ml molten top agar (0.5%), which was then poured over a prewarmed LB agar plate. The plates were incubated overnight at 37 °C and the plaques generated by temperate and virulent phages were discriminated by differences in opacity, and a subset confirmed by PCR. All experiments were performed in six replicates.

**Coevolution competitions.** Phage competition was also measured in the presence of host evolution. Initially sensitive  $\Delta cas7$  or wild-type hosts were infected with 10<sup>4</sup> PFU of a 50:50 mix of temperate and virulent phages (DMS3 and DMS3vir). The experiment was run as a standard evolution experiment, and bacterial and phage titres measured every day for seven days. Resistance phenotypes were assessed on days 3 and 7, as described in 'Determination of resistance phenotypes', and plaque assays used to determine the ratio of temperate to virulent phage at each time point.

**Competition DMS3 × DMS3 *acrIF*.** Temperate DMS3 was also competed against a mutant encoding Acr proteins. Initially sensitive  $\Delta cas7$  or wild-type hosts were infected with 10<sup>4</sup> PFU of a 50:50 mix of DMS3 and DMS3 *acrIF1* (or DMS3 *acrIF4*). The experiment was run for three days and samples of total phage population (that is, free phages + prophages, no chloroform extraction) were collected and immediately frozen every day. Phage titres were measured every day by spot test. To determine their relative fitness, the relative frequencies of each phage were determined at  $t = 0$  and at  $t = 3$  days by qPCR, following a previously described method<sup>32</sup>.

### Expression of priming spacers

The expression of the CRISPR RNA encoding spacer 1 of CRISPR2 was restored in strain PA14  $\Delta CRISPR2$  using an arabinose inducible expression vector (pHERD30T-based). In brief, oligonucleotides containing the sequences of a non-targeting (NT) spacer (5'-GTCTTCTTTGAGCTTCCAGAGAACTGAAGAC-3') or of spacer 1 (5'-ATCAGCCGGACGTGTAGTAGTCGAGCGCGGT-3') and flanked by two CRISPR2 repeats were annealed and ligated between NcoI/HindIII restriction sites. The resulting plasmids were transformed into PA14  $\Delta CRISPR2$  to generate the strains  $\Delta CRISPR2$ -sp1 and  $\Delta CRISPR2$ -NT, respectively. Similarly, a spacer targeting the natural Pf5 prophage of PA14 (accession number AY324828) with 1 mismatch (bold) 5'-AGTCCTTCTAGTGACGGAA CCAAATCTATT-3' was expressed in the wild-type PA14 strain.

### Measuring prophage induction

Single colonies were picked from plated lysogens and grown in LB medium overnight at 37 °C with 180 rpm shaking. The overnight culture

was diluted 1:100 in fresh M9 medium and grown until OD<sub>600nm</sub> reached about 0.1. The cultures were then pelleted by centrifugation and the pellet washed 5 times in M9 buffer, and then resuspended in 10 ml of M9 medium. The cultures were grown at 37 °C with shaking, samples were taken regularly for phage quantification by spot assay and OD<sub>600nm</sub> measurements using the Biotek synergy 2 plate reader.

### Twenty-four-hour growth curves

Single colonies were isolated and grown in M9 medium overnight at 37 °C with 180 rpm shaking. The following day, this culture was diluted 1:100 and 250 µl of this mixture was added to a 96-well plate and growth curves were measured for 23 h in a Thermo Scientific Varioskan flash plate reader with continuous shaking at 180 rpm. Readings of OD<sub>600nm</sub> were taken every 15 min and the plate kept at 37 °C. All growth curves were performed in 6–12 replicates.

### Bacterial competition

Competition experiments were performed in 6 ml M9 medium supplemented with 0.2% glucose. Competition experiments were initiated by inoculating 1:100 from a 1:1 mixture of overnight cultures (grown in M9 medium + 0.2% glucose) of each strain. If phages were included, they were added at a concentration of 10<sup>4</sup> PFU. Cells were transferred 1:100 daily into fresh broth. At varying time points, from 1 to 7 days, samples were taken and cells were serially diluted in M9 salts and plated on LB agar supplemented with 50 µg ml<sup>-1</sup> X-gal (to enable discrimination between strains that carry the *lacZ* gene (blue) and those that do not (white)). All experiments were performed in six replicates. Relative fitness was calculated from changes in the relative frequencies of blue and white colonies (relative fitness = [(fraction strain A at  $t = x$ ) × (1 – (fraction strain A at  $t = 0$ ))]/[(fraction strain A at  $t = 0$ ) × (1 – (fraction strain A at  $t = x$ ))]).

### Whole-genome sequencing and bioinformatic analyses

Lysogen clones in the *Δcas7* background and lysogen clones in wild-type and *Δcas1* backgrounds were isolated from late time points of the evolution experiments. Standard genome sequencing and standard bioinformatic analyses were provided by MicrobesNG (as described at <http://www.microbesng.uk>). Raw read sequencing data has been deposited in the European Nucleotide Archive under the study accession number PRJEB34503. Trimmed reads were mapped to wild-type PA14 reference genome (accession number NC\_008463) with Geneious 9.1.8 software using Bowtie2 mapper<sup>33</sup> to identify the genomic deletion. Reads were also mapped to DMS3 reference genome (accession number DQ631426) and hybrid reads composed of a 5' extremity deriving from PA14 and a 3' extremity matching the DMS3 5' end (and vice versa, read 5' extremity matching DMS3 3' end and 3' extremity deriving from PA14) were extracted. These hybrid reads were then mapped back to PA14 genome to identify prophage insertion sites.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

Source Data associated with Figs. 1–4 and Extended Data Figs. 1–3, 5, 7–9 are provided with the paper. Sequencing data have been deposited in the European Nucleotide Archive under the study accession number PRJEB34503. The datasets analysed for the bioinformatic study are available on GitHub at <https://github.com/davidchyou/Rollie-Chevallereau>.

### Code availability

Mathematical algorithms generated during this study are available in the Supplementary Information. Scripts generated for the bioinformatics analyses are available on GitHub at <https://github.com/davidchyou/Rollie-Chevallereau>.

28. Cady, K. C. & O'Toole, G. A. Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *J. Bacteriol.* **193**, 3433–3445 (2011).
29. Liberati, N. T. et al. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl Acad. Sci. USA* **103**, 2833–2838 (2006).
30. Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol.* **194**, 5728–5738 (2012).
31. Ceyssens, P.-J. et al. Comparative analysis of the widespread and conserved PB1-like viruses infecting *Pseudomonas aeruginosa*. *Environ. Microbiol.* **11**, 2874–2883 (2009).
32. Chevallereau, A. et al. Exploitation of the cooperative behaviors of anti-CRISPR phages. *Cell Host Microbe* **27**, 1–10 (2019).
33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
34. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).

**Acknowledgements** We thank A. R. Davidson for providing the mutant strains of PA14 *Δcas3*, *Δcas7*, *Δcas1* and *ΔCRISPR2*, G. A. O'Toole for the strain CRISPR2 *Δsp1-2* and J. Bondy-Denomy for the *Tn::pilA* (*ΔpilA*) PA14 surface mutant. Genome sequencing was provided by MicrobesNG (<http://www.microbesng.uk>), which is supported by the BBSRC (grant number BB/L024209/1). This work was funded by a grant from the European Research Council (<https://erc.europa.eu>) (ERC-STG-2016-714478 - EVOIMMECH) and NERC Independent Research Fellowship (NE/M018350/1) awarded to E.R.W. A.C. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 834052. P.C.F. was supported by the Marsden Fund from the Royal Society of New Zealand.

**Author contributions** Conceptualization of the study was done by C.R., A.C. and E.R.W. Experimental design was carried out by C.R., A.C., B.N.J.W. and E.R.W. Bacterial evolution, competition and growth experiments were done by C.R. and A.C. with assistance from O.F. and I.M. Virulent versus temperate phage competitions and prophage induction-rate experiments were performed by C.R. All experiments with *acr* phages and CRISPR 2 spacer 1 were done by A.C. B.N.J.W. and A.C. carried out Pf5 experiments. The experiment with superinfecting virulent phage was done by E.R.W. C.R., A.C., B.N.J.W., T.-y.C., C.M.B., P.C.F. and E.R.W. analysed the data. S.G. generated theoretical mathematical models. T.-y.C. conducted bioinformatic analyses supervised by C.M.B. and P.C.F. A.C. performed whole-genome sequencing analyses. C.R. wrote the original draft of the manuscript; A.C. wrote the revised version of the manuscript with contributions from C.R., B.N.J.W., T.-y.C., S.G., C.M.B. and P.C.F. E.R.W. supervised the project and provided comments on all versions of the manuscript.

**Competing interests** The authors declare no competing interests.

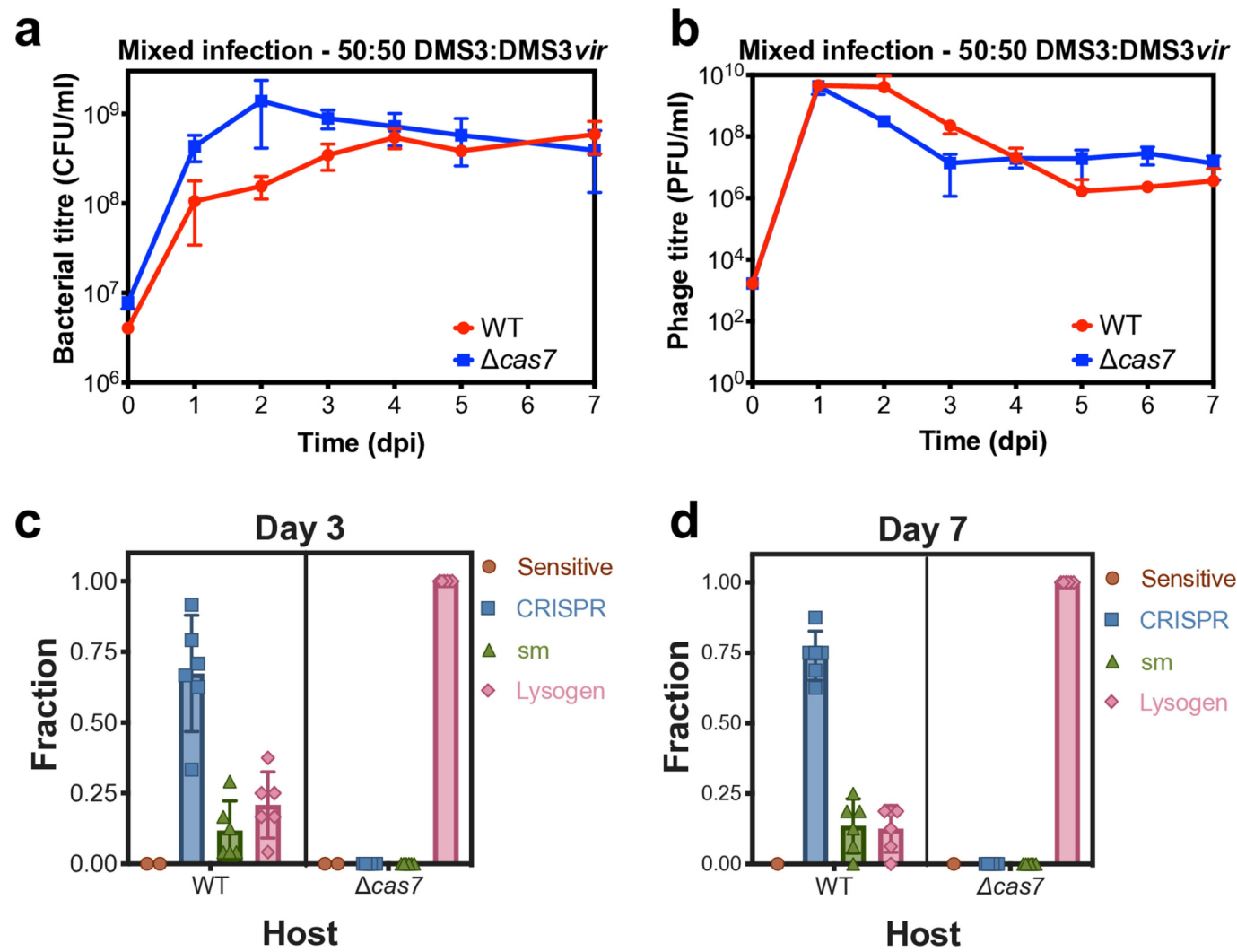
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1936-2>.

**Correspondence** and requests for materials should be addressed to C.R., A.C. or E.R.W.

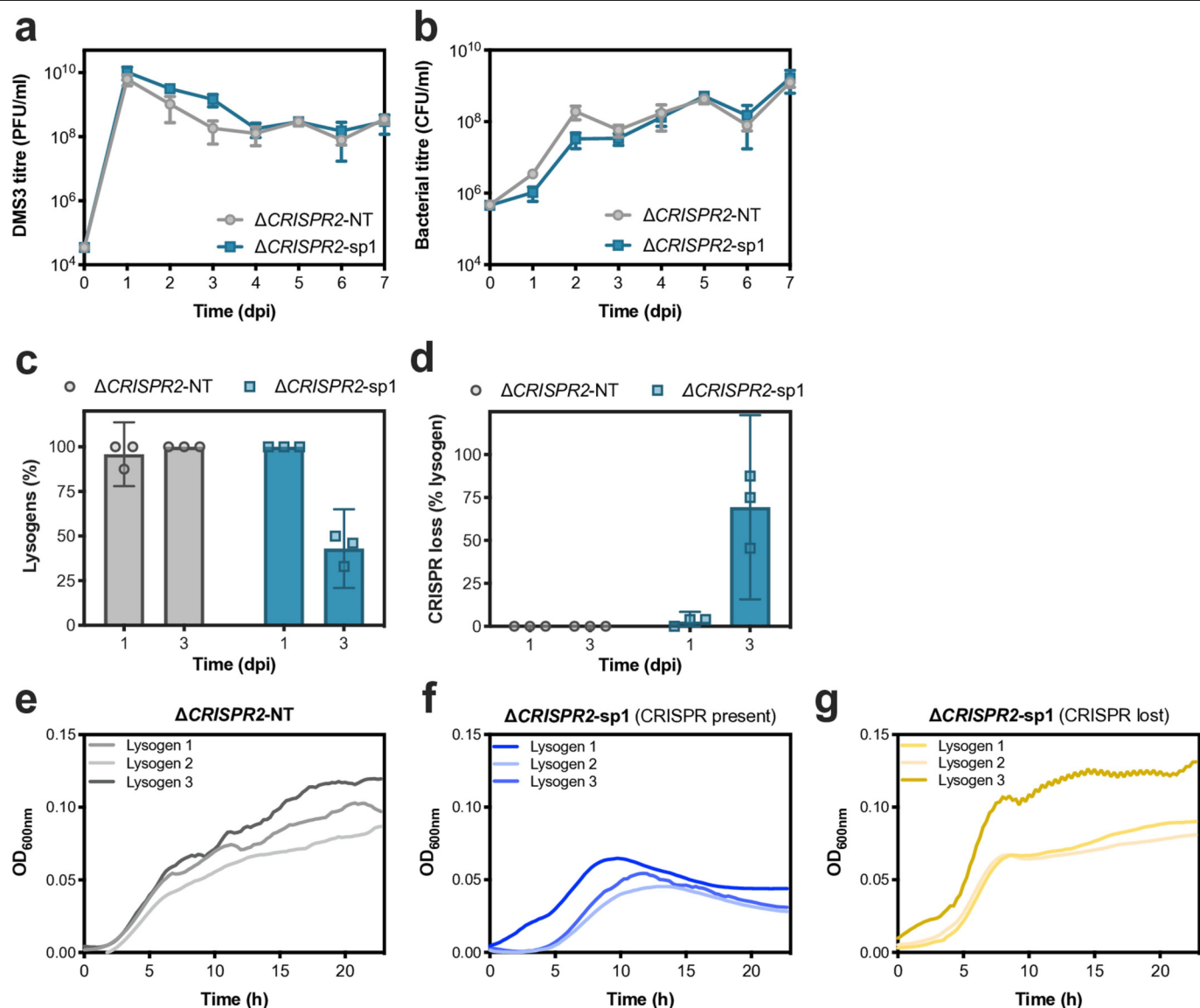
**Peer review information** Nature thanks Eugene Koonin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



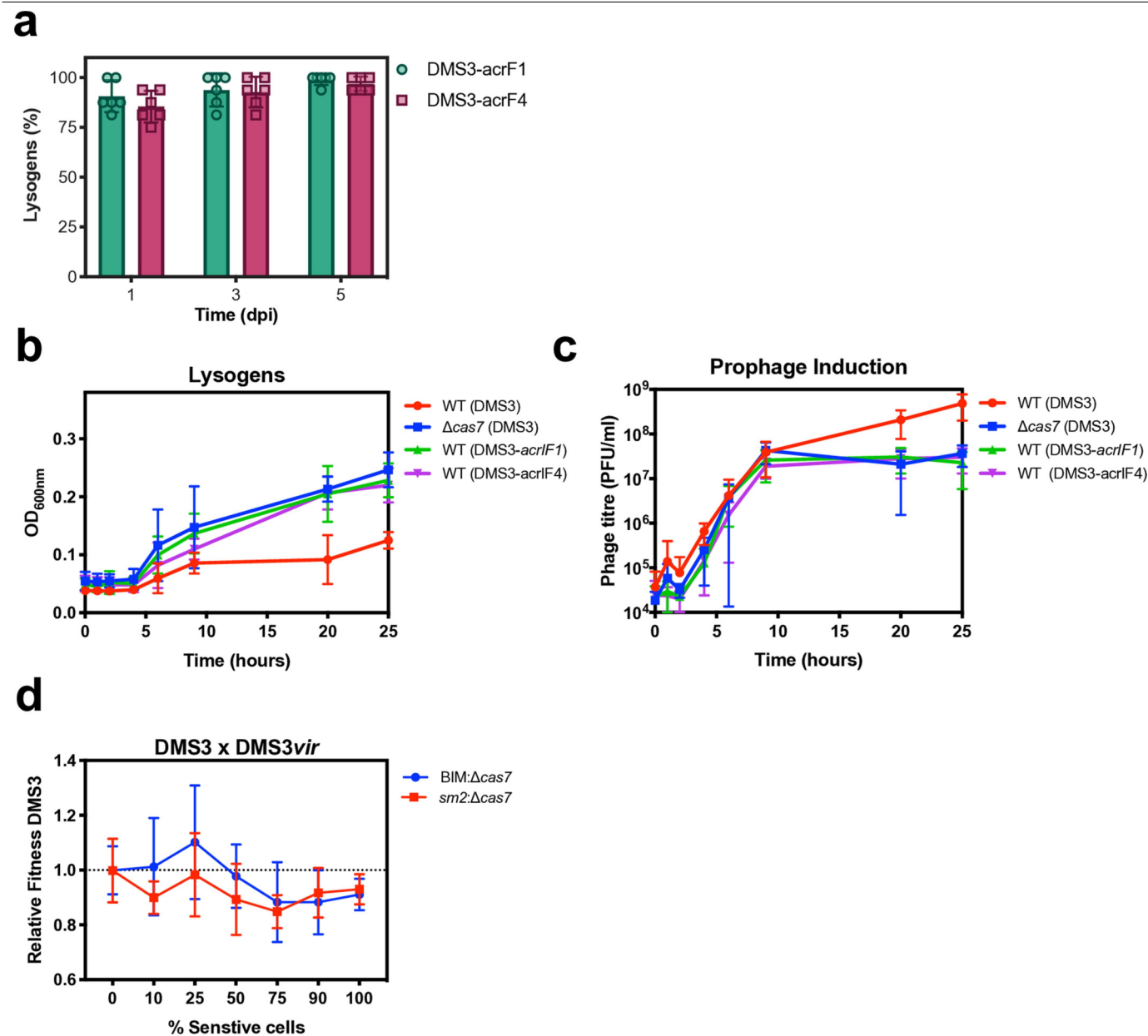
**Extended Data Fig. 1 | Infection with a 50:50 mix of temperate:virulent phages. a, b,** Bacterial (a) and phage (b) titres during a co-culture experiment of wild-type PA14 (red) or  $\Delta cas7$  mutant (blue), and a 50:50 mix of DMS3 and DMS3vir. **c, d,** Resistance phenotypes at day 3 (c) or day 7 (d) of the co-culture

experiment, based on 24 random clones per replicate experiment. Data are the mean of six biological replicates per treatment. Error bars represent 95% confidence intervals.



**Extended Data Fig. 2 | The suppression of lysogeny and immunopathological effects are due to spacer 1 of CRISPR array 2.** **a, b**, Phage (a) and bacterial (b) titres during co-culture of phage DMS3 and *P. aeruginosa* PA14  $\Delta$ CRISPR2, expressing a non-targeting spacer from a plasmid ( $\Delta$ CRISPR2-NT) or the original CRISPR2 spacer 1 ( $\Delta$ CRISPR2-sp1). **c, d**, The proportion of lysogens (c) and the frequency of loss of CRISPR–Cas immune systems (d) at 1 and 3 days post-infection, based on PCR analyses of 24 random clones per replicate

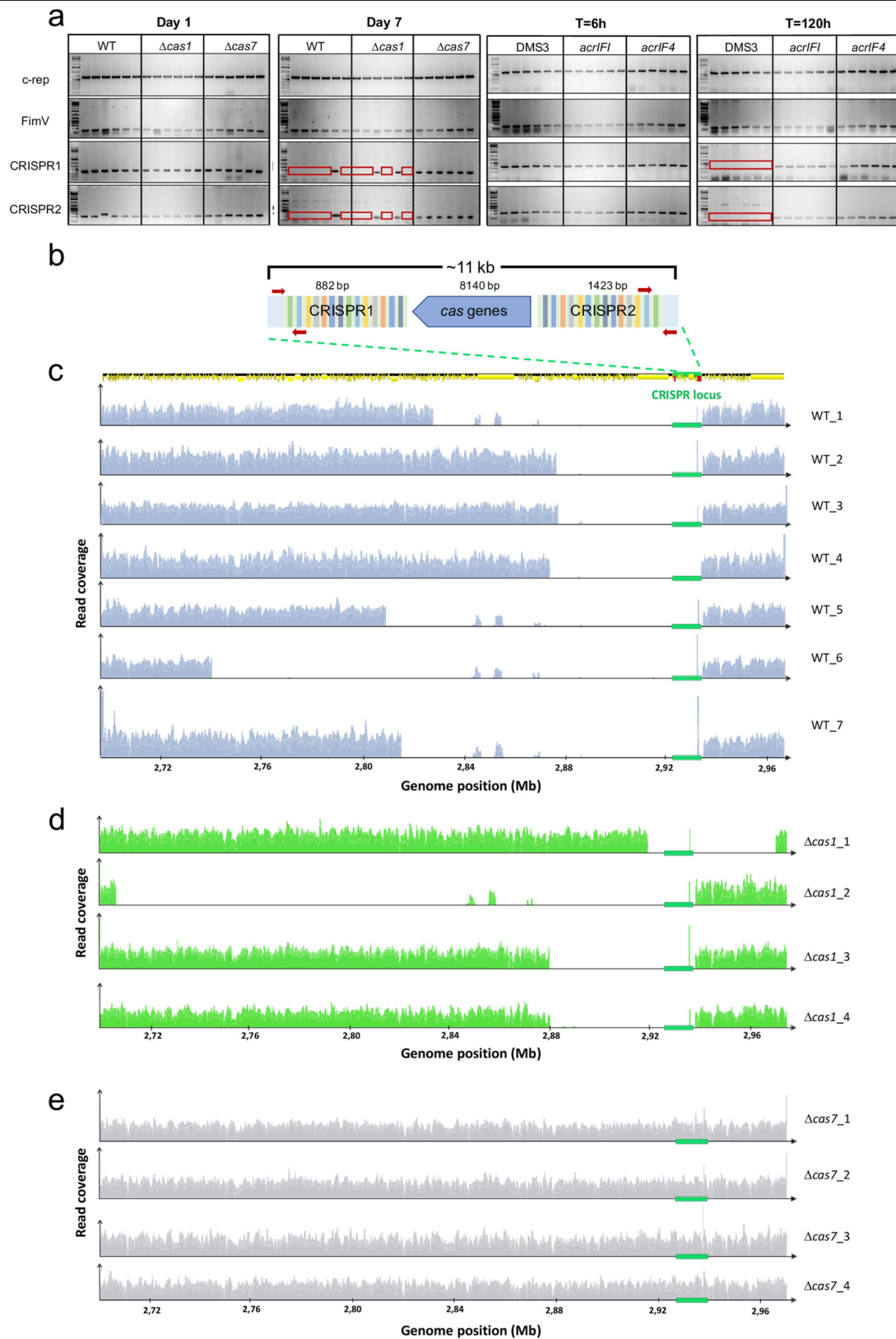
experiment. **a–d**, Data are the mean of three biological replicates. Error bars represent 95% confidence intervals. **e–g**, Growth of three independent lysogen clones isolated at three days post-infection, as determined by OD<sub>600nm</sub> measurements.  $\Delta$ CRISPR2-NT (e) and  $\Delta$ CRISPR2-sp1 (f) lysogen clones carry the ancestral  $\Delta$ CRISPR2 CRISPR–Cas immune system, whereas the  $\Delta$ CRISPR2-sp1 (g) lysogen clones have evolved to lose CRISPR–Cas.



**Extended Data Fig. 3 | Prophage induction rates are increased in hosts with active CRISPR-Cas.** **a**, The percentage of lysogens formed upon infection of wild-type host with DMS3 phages engineered to produce AcrIF1 or AcrIF4 anti-CRISPR proteins. **b**, **c**, Optical density (**b**) and phage titres (**c**) during growth of lysogens of DMS3, DMS3 *acrIF1* or DMS3 *acrIF4* phages in a wild-type PA14 or  $\Delta cas7$  genetic background. **d**, Relative fitness of the DMS3 phage during

competition with the virulent mutant DMS3vir in the presence of varying fractions of sensitive ( $\Delta cas7$ ) host and resistant hosts with CRISPR-based immunity (BIM) or surface-based immunity (sm2) against these phages. Data show mean fitness at 8 h after infection. All panels show the mean of six biological replicates and error bars represent 95% confidence intervals.



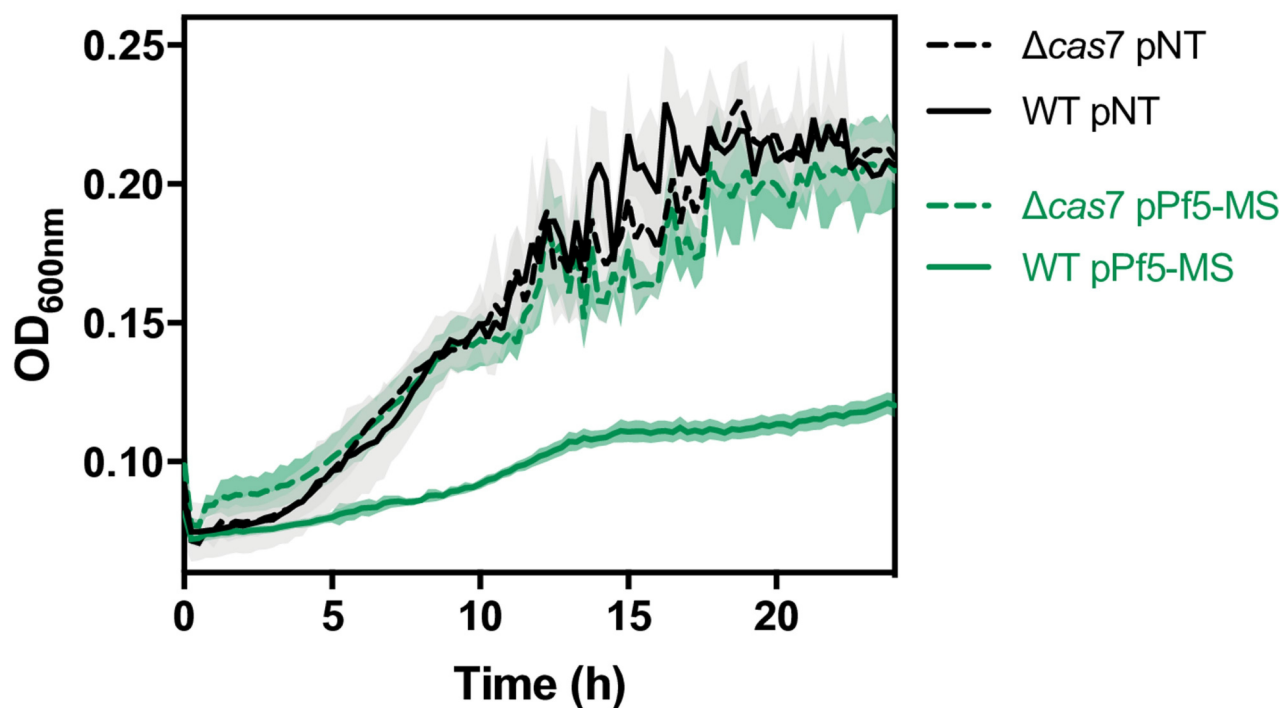
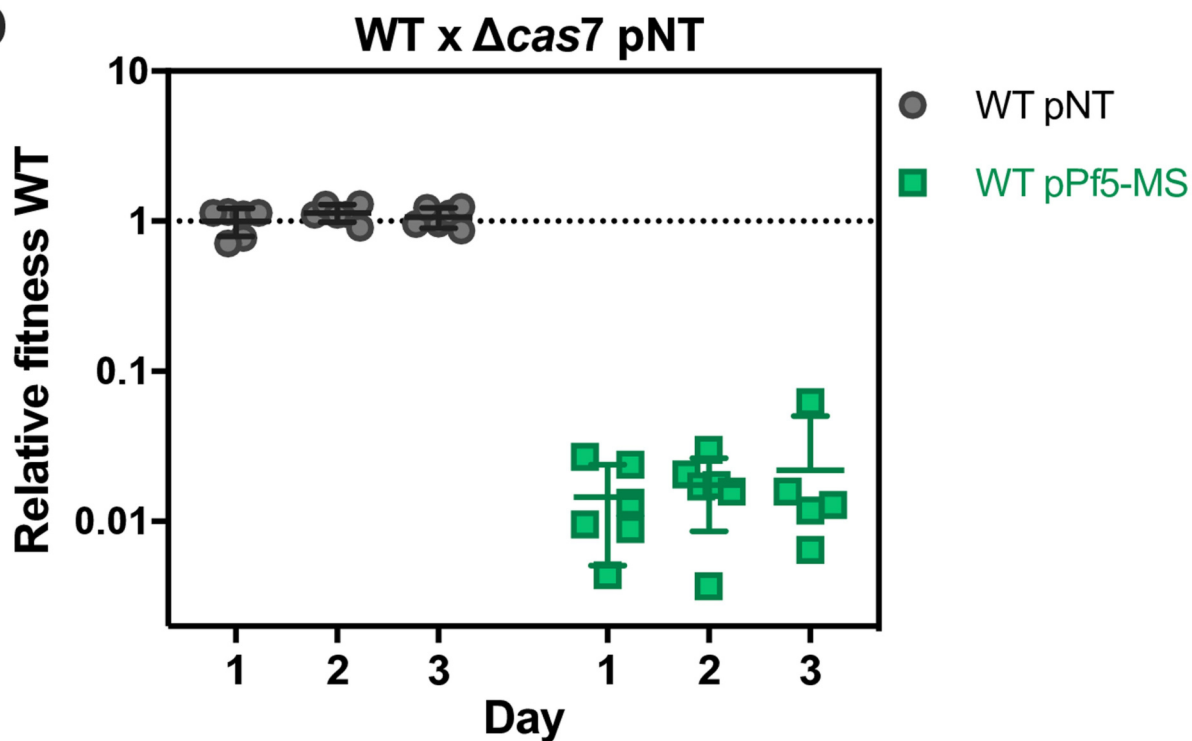


**Extended Data Fig. 4** | See next page for caption.

## Extended Data Fig. 4 | Lysogens lose their CRISPR–Cas immune systems.

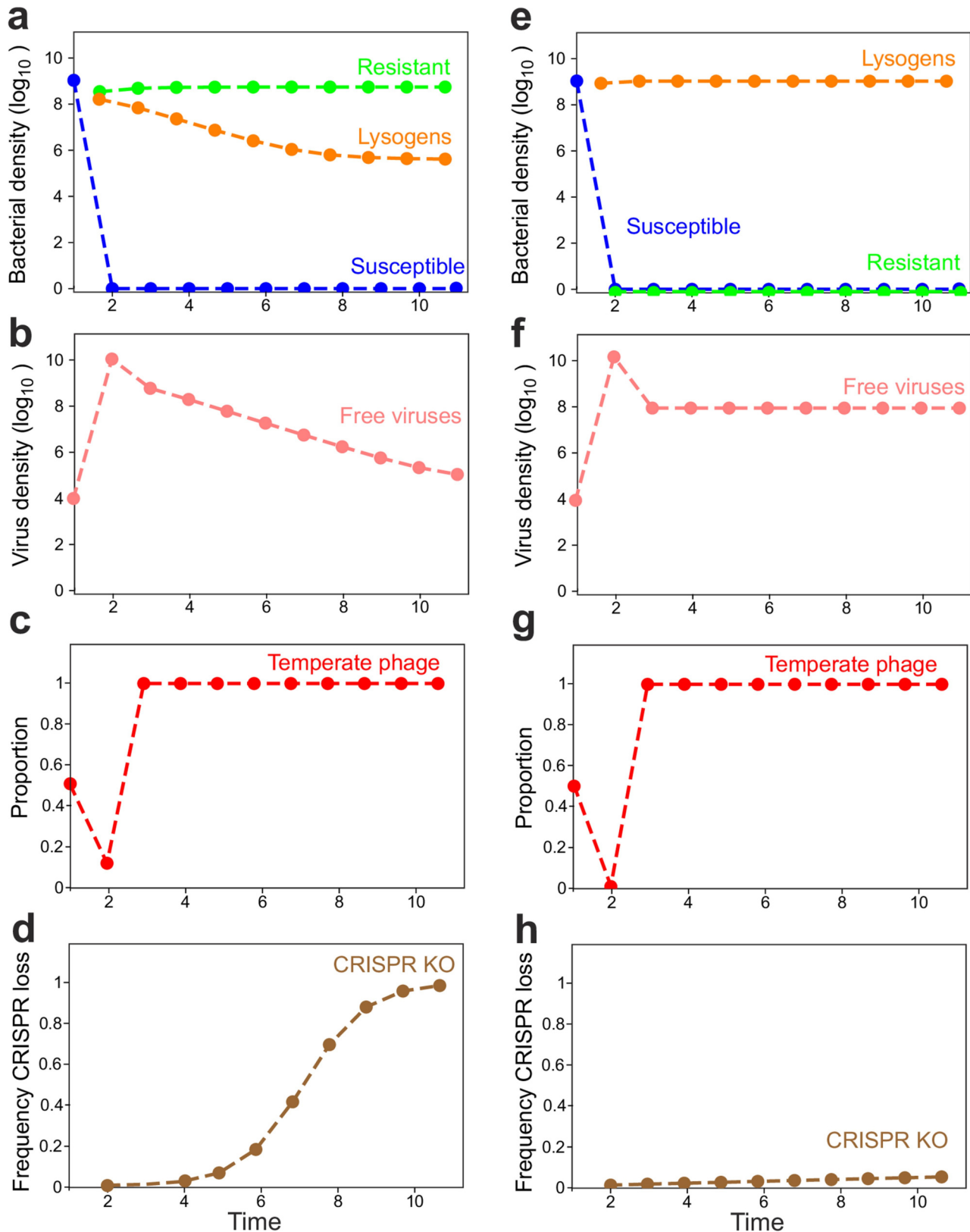
**a**, PCR amplification of the *c*-repressor gene of the prophage (*c-rep*, 611 bp), the *fimV* gene (located about 1 Mb from the CRISPR loci and used as a positive control for the PCR, 116 bp) and CRISPR loci 1 (349 bp) and 2 (206 bp) on the host genome. PCRs were performed on 6 independent DMS3 lysogens in wild-type,  $\Delta cas1$  and  $\Delta cas7$  backgrounds isolated at 1 or 7 days post-infection, as well as on 6 independent lysogens of DMS3, DMS3 *acrIF1* or DMS3 *acrIF4* (wild-type background) isolated at 6 or 120 h after infection. Red frames indicate a failure to amplify a product. PCR amplifications were performed on clones isolated from three biological replicate experiments and produced similar results. For gel source data, see Supplementary Fig. 1. **b**, Schematic of the CRISPR–Cas locus of wild-type PA14, which spans a region of around 11 kb. Primers used to amplify regions of CRISPR arrays 1 or 2 are shown as red arrows. **c–e**, Whole-

genome sequencing of DMS3 lysogens that lost their CRISPR–Cas system (red frames in **a**) in wild-type PA14 (**c**),  $\Delta cas1$  (**d**) or  $\Delta cas7$  (**e**) backgrounds. Graphs show the read coverage of the region encompassing positions 2.70–2.97 Mb of the wild-type PA14 genome. The CRISPR–Cas locus is indicated by a green box on the x axis. A genome map depicting coding sequences (yellow arrows) is shown above the graphs. The region comprising 2.84–2.88 Mb includes sequences that are repeated elsewhere on the PA14 genome, explaining why reads that map to these positions are still detected in some of the deletion mutants. The high peak at the 3' end of the CRISPR locus corresponds to the coverage of spacer 20 of CRISPR2 by reads that derive from DMS3 prophage (5' and 3' extremities of these reads map to the phage genome). Spacer 20 of CRISPR2 has 100% identity to DMS3 but is not immunogenic because there is no consensus protospacer-adjacent motif.

**a****b**

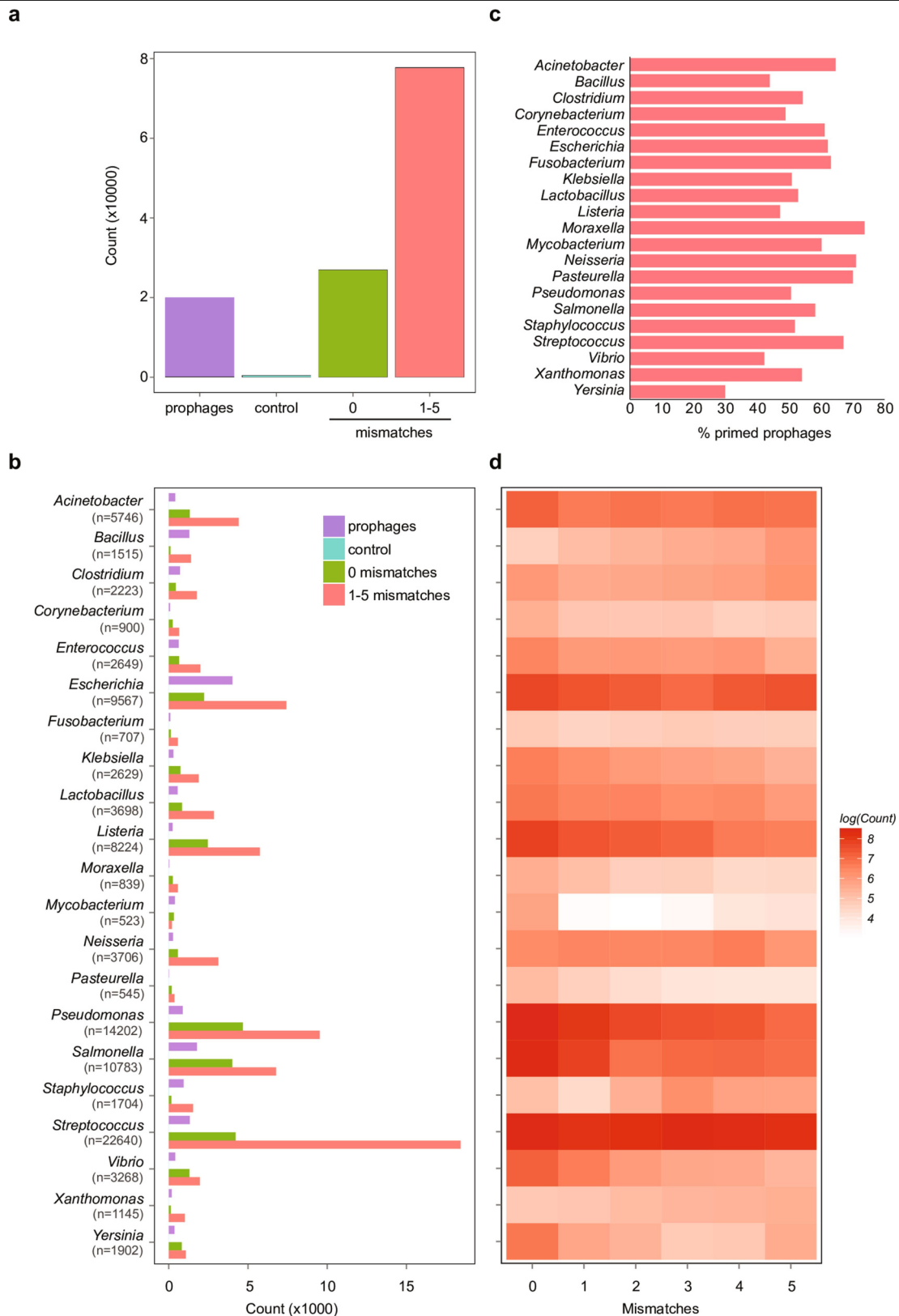
**Extended Data Fig. 5 | Expression of Pf5 priming spacer in *P. aeruginosa* PA14.** **a**, Growth of  $\Delta cas7$  (dashed line) or wild-type (solid line) clones carrying an expression plasmid encoding a non-targeting spacer (pNT) or a spacer targeting the PA14 natural prophage Pf5 with one mismatch (pPf5-MS), as determined by OD<sub>600nm</sub> measurements. Graphs show mean curves from 6

biological replicates, and shaded areas correspond to 95% confidence intervals. **b**, Relative fitness of wild-type pNT or wild-type pPf5-MS during competition with  $\Delta cas7$  pNT. Data are the mean of six biological replicates per treatment. Error bars represent 95% confidence intervals.



**Extended Data Fig. 6 | Simulations of population and evolutionary dynamics of bacteria-phage interactions, when virulent and temperate phages compete on bacteria with a CRISPR-Cas system. a–c, e–g.** Graphs show densities of susceptible hosts, CRISPR-resistant bacteria and lysogens (a, e) or free viruses over time (b, f), as well as the proportion of temperate phages in a population composed of both temperate and virulent types (c, g). Temperate

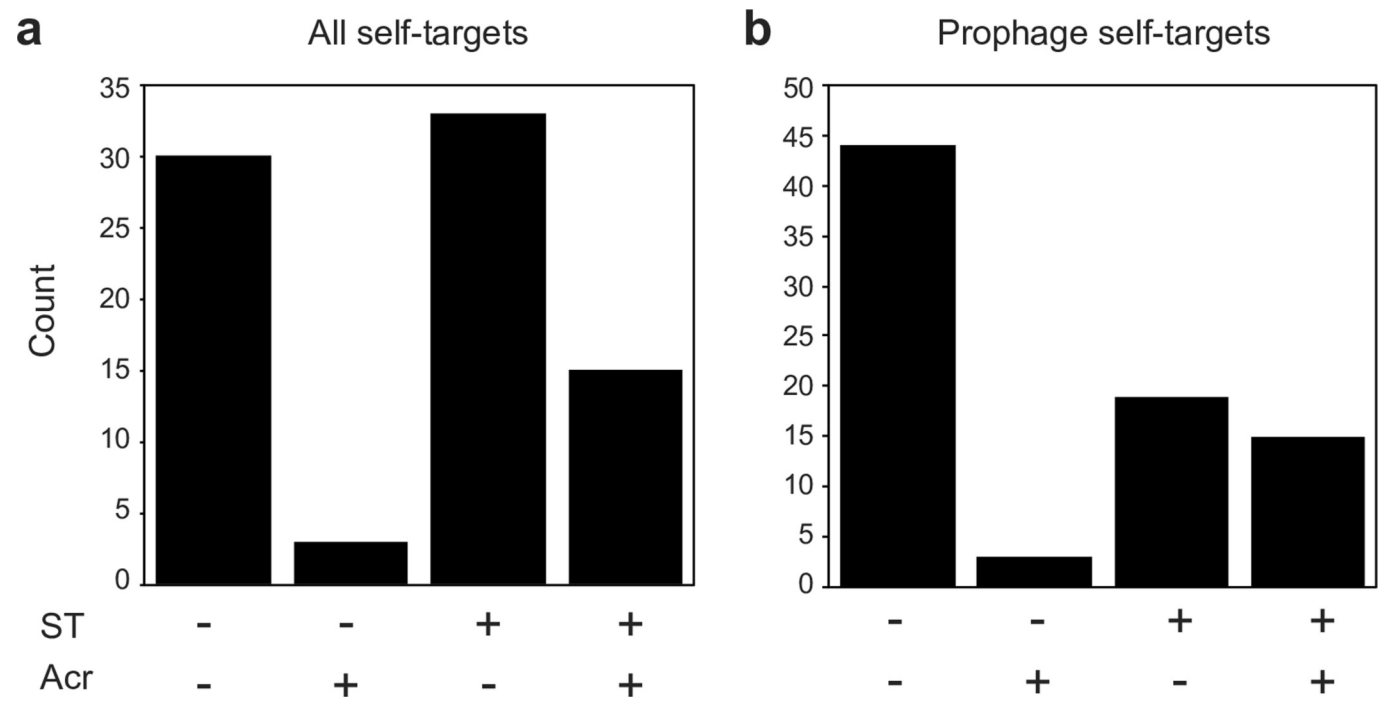
phages can transmit both horizontally and vertically, whereas virulent phages can transmit only horizontally and cannot superinfect lysogens. d, h. Frequency of evolutionary loss of CRISPR-Cas system in the lysogen population over time. The simulations shown in a–d reflect a situation in which both virulent and temperate phages lack *acr* genes, whereas those in e–h reflect a scenario in which the temperate type carries an *acr* gene.



**Extended Data Fig. 7 | Matches between spacers and temperate phages are widespread.** **a**, Total matches between non-redundant spacers ( $n = 1,239,973$ ) from 171,361 RefSeq and GenBank complete genomes and a non-redundant set of temperate phages ( $n = 19,996$ )<sup>21</sup>. The counts of perfect (0) or mismatched (1–5) targets are shown. As a control, the temperate phages were shuffled ten times, while retaining the hexanucleotide content (control). **b**, Counts of spacers matching temperate phages from all genera with over 500 spacer–phage matches. The total number ( $n$ ) of spacer–phage matches is shown

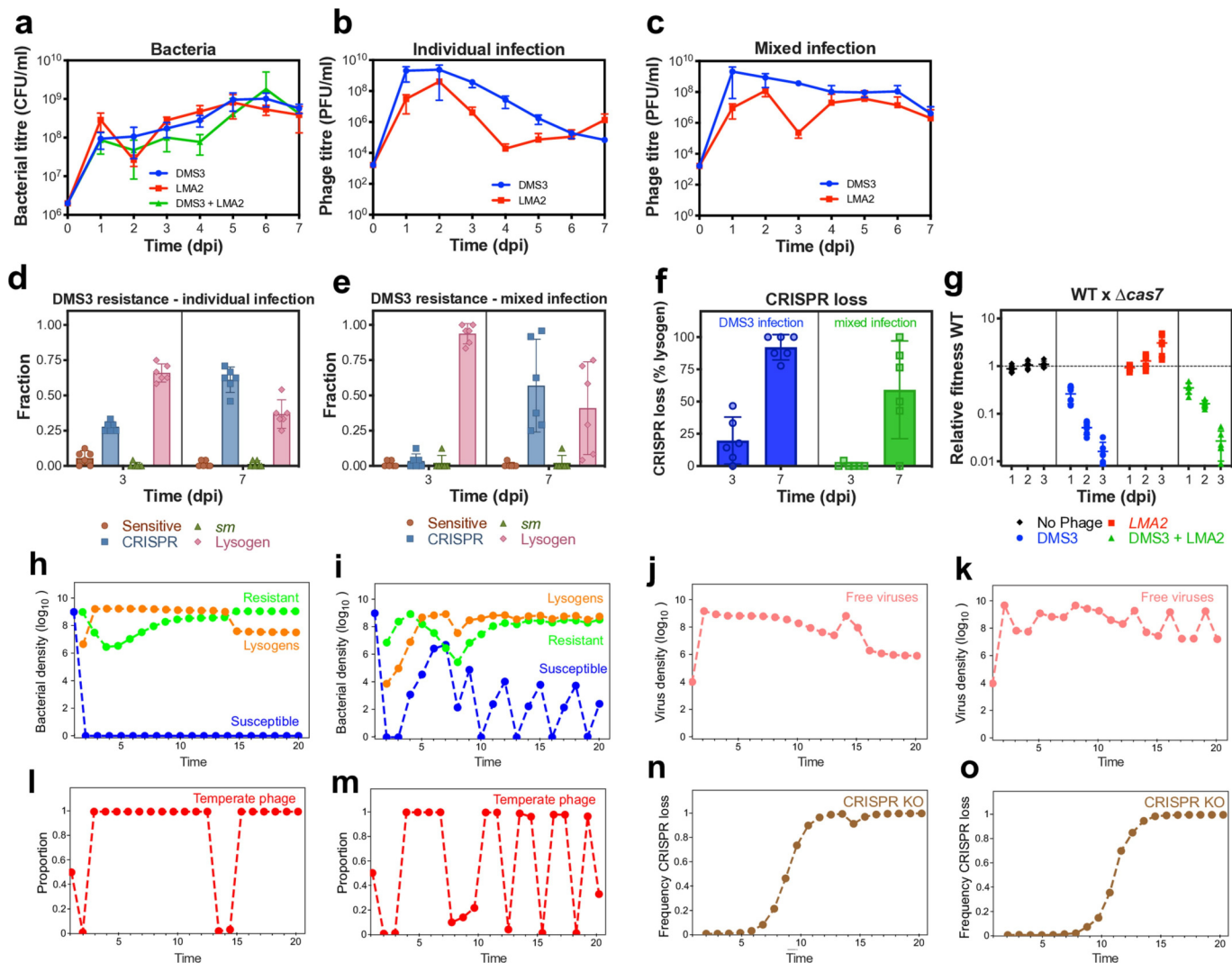
for each genus in parentheses. Counts of matches are shown (0, green; 1–5 mismatches, red). The number of temperate phages analysed is plotted (prophages in purple) as are the matches to shuffled prophages. The control is shown in blue, but is not visible because it had only 0 to 10 counts. **c**, The percentage of prophages within each genus that were targeted by self-priming spacers (1–5 mismatches). **d**, Heat map of the distribution of mismatches (0–5). Genera are as in **b** and data are shown as  $\log(\text{count})$  for each genus, as the number of matches varied widely between genera.





**Extended Data Fig. 8 | Self-targeting genomes are enriched for *acr* gene(s).**  
**a, b,** The number of *P. aeruginosa* genomes with complete CRISPR–Cas systems that contain (+) or lack (–) genes encoding known Acr proteins. For these strains, the total number of strains with perfect (0) or mismatched (1–5) self-targeting (ST) spacers to anywhere in the genome (**a**) or to prophages (**b**) are

shown. For complete *P. aeruginosa* genomes, all self-targeting events were analysed for matches to prophages using PHASTER<sup>34</sup>. The number of genomes with *acr* genes (*acr* +) and self-targeting (ST +) spacers is significantly greater than the number of genomes with *acr* genes and without self-targeting spacers ( $P=8.14 \times 10^{-5}$ , two-sided Fisher’s exact test,  $n=71$ ).



**Extended Data Fig. 9 | Presence of a superinfecting virulent phage does not alter immunopathological effects.** **a–c**, Bacterial (**a**) and phage titres upon individual (**b**) or mixed (**c**) infection of wild-type PA14 with phage DMS3 and virulent phage LMA2. **d, e**, Resistance phenotypes evolved by bacteria against DMS3 upon individual (**d**) or mixed (**e**) infection. **f**, Frequency of loss of CRISPR–Cas immune systems upon infection with phage DMS3 or with both the phages DMS3 and LMA2, based on 24 random clones per replicate experiment. **g**, Relative fitness of wild-type PA14 during competition with PA14  $\Delta cas7$  in the presence or absence of phages DMS3 and LMA2. **a–g**, Data are the means of six biological replicates. Error bars indicate 95% confidence intervals.

**h–o**, Simulations of population and evolutionary dynamics during infection of bacteria carrying CRISPR–Cas systems with a mixed population of unrelated virulent and temperate phages. Graphs show densities of susceptible hosts,

CRISPR-resistant bacteria and lysogens (**h, i**) and free viruses over time (**j, k**), as well as the frequencies of temperate phages in a population composed of both temperate and virulent types (**l, m**). Temperate phage can transmit both horizontally and vertically, whereas virulent phage can transmit only horizontally and can superinfect the lysogens (because temperate and virulent phages are unrelated). **n, o**, Frequencies of evolutionary loss of CRISPR–Cas system in the lysogen population over time. The simulations shown in **h, j, l, n** reflect a scenario in which bacteria can evolve CRISPR-based resistance against both phages, whereas those shown in **i, k, m, o** reflect a situation in which CRISPR-based resistance does not evolve against the virulent phage, and bacteria instead evolve costly surface-based resistance (as it is the case in our experiments). A detailed description of the simulations is provided in the Supplementary Information.

Extended Data Table 1 | Genomic deletions and prophage insertion sites in DMS3 late lysogen clones

Strain	Sample name	Name on ED Fig. 4	Deleted region			Prophage insertion site(s)	
			Start (bp)	End (bp)	Length (bp)	Position (bp)	Proportion of mapped hybrid reads <sup>a</sup>
PA14 WT	WT117	WT_1	2,831,115	2,938,245	<b>107,131</b>	replace deleted region	73
	WT221	WT_2	2,879,871	2,938,280	<b>58,410</b>	replace deleted region	44.2
						3,555,252	42.3
	WT3213	WT_3	2,880,477	2,938,285	<b>57,809</b>	replace deleted region	84
						replace deleted region	72.3
						3,944,997	6.3
	WT4313	WT_4	2,878,050	2,938,509	<b>60,460</b>	4,067,332	3.8
						4,453,953	2.5
	WT566 <sup>b</sup>	-	0	0	<b>0</b>	nd	nd
	WT6615	WT_5	2,812,027	2,938,165	<b>126,139</b>	replace deleted region	42.0
						3,325,229	49.6
						replace deleted region	22.6
	WT7AnneD1	WT_6	2,743,638	2,938,176	<b>194,539</b>	5,214,542	28.9
						5,732,030	24.0
						5,860,538	22.6
PA14 $\Delta$ cas1						2,335,454	24.8
						2,652,167	29.6
	WT8AnneD2	WT_7	2,818,271	2,938,290	<b>120,020</b>	replace deleted region	21.8
						3,079,052	21.3
	cas11114	$\Delta$ cas1_1	2,918,943	2,970,427	<b>51,485</b>	3,079,190	84.3
	cas1222	$\Delta$ cas1_2	2,706,504	2,938,538	<b>232,035</b>	replace deleted region	90.6
						4,478,625	2.8
	cas1333 <sup>c</sup>	-	0	0	<b>0</b>	partial	nd
PA14 $\Delta$ cas7						2,730,814	44.1
	cas14415	$\Delta$ cas1_3	2,879,882	2,938,523	<b>58,642</b>	replace deleted region	46.0
						4,481,869	3.1
	cas1555 <sup>d</sup>	-	2,902,848	2,938,280	<b>35,433</b>	replace deleted region	78.8
						2,934,176	7.7
	cas16611	$\Delta$ cas1_4	2,879,874	2,938,285	<b>58,412</b>	replace deleted region	94.2
PA14 $\Delta$ cas7	cas7114	$\Delta$ cas7_1	0	0	<b>0</b>	1,120,809	86.7
	cas72211	$\Delta$ cas7_2	0	0	<b>0</b>	410,059	82.9
	cas73313	$\Delta$ cas7_3	0	0	<b>0</b>	5,725,887	74.7
	cas7442	$\Delta$ cas7_4	0	0	<b>0</b>	5,743,045	84.5
	cas7554	-	0	0	<b>0</b>	905,180	37.9
	cas7669	-	0	0	<b>0</b>	1,667,097	75.0

<sup>a</sup>The sum does not reach 100% because a proportion of hybrid reads that mapped to the PA14 genome did not enable the identification of a potential prophage insertion site.

<sup>b</sup>Sample was contaminated with a BIM clone; sequencing data are not interpretable.

<sup>c</sup>Only 2,607 bp of DMS3 DNA (including the c-repressor gene) are inserted in bacterial genome at position 5,834,730.

<sup>d</sup>A mixed population, in which reads matching the CRISPR locus are still detectable but coverage is very low.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection

Data analysis

Data analysis and figures were made in R version 3.5.1. or Graphpad Prism 7. Whole-genome sequences analyses were done in Geneious® 9.1.8 software using Bowtie2 mapper. Custom scripts used in bioinformatics analyses are available on github at <https://github.com/davidchyou/Rolie-Chevallereau>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Source data associated with main figures and Extended Data Figures 1,2,3,5,7,8 and 9 are available in the online version of this publication. Sequencing data have been deposited in the European Nucleotide Archive under the study accession number PRJEB34503. The datasets analysed for the bioinformatic study are available on github at <https://github.com/davidchyou/Rolie-Chevallereau>

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences      ☐ Behavioural & social sciences      ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All experiments were performed in 6 independent biological replicates for sufficient statistical power, in line with common practice in experimental evolution studies.
Data exclusions	No data were excluded
Replication	We used 6 independent biological replicates per treatment. All observations were reproducible.
Randomization	Not relevant to our study - single experimental manipulations of individual variables.
Blinding	Blinding not relevant - single experimental manipulations of individual variables.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging



# Robust and persistent reactivation of SIV and HIV by N-803 and depletion of CD8<sup>+</sup> cells

<https://doi.org/10.1038/s41586-020-1946-0>

Received: 12 March 2019

Accepted: 12 December 2019

Published online: 22 January 2020

There are amendments to this paper

Julia Bergild McBrien<sup>1</sup>, Maud Mavigner<sup>2</sup>, Lavinia Franchitti<sup>1</sup>, S. Abigail Smith<sup>1</sup>, Erick White<sup>1</sup>, Gregory K. Tharp<sup>1</sup>, Hasse Walum<sup>1</sup>, Kathleen Busman-Sahay<sup>3</sup>, Christian R. Aguilera-Sandoval<sup>4</sup>, William O. Thayer<sup>4</sup>, Rae Ann Spagnuolo<sup>4</sup>, Martina Kovarova<sup>4</sup>, Angela Wahl<sup>4</sup>, Barbara Cervasi<sup>1</sup>, David M. Margolis<sup>4,5</sup>, Thomas H. Vanderford<sup>1</sup>, Diane G. Carnathan<sup>1</sup>, Mirko Paiardini<sup>1,6</sup>, Jeffrey D. Lifson<sup>7</sup>, John H. Lee<sup>8</sup>, Jeffrey T. Safrin<sup>8</sup>, Steven E. Bosinger<sup>1,6</sup>, Jacob D. Estes<sup>3,9</sup>, Cynthia A. Derdeyn<sup>1,6</sup>, J. Victor Garcia<sup>4</sup>, Deanna A. Kulpa<sup>1,6</sup>, Ann Chahroudi<sup>2,10,11</sup> & Guido Silvestri<sup>1,6,11\*</sup>

Human immunodeficiency virus (HIV) persists indefinitely in individuals with HIV who receive antiretroviral therapy (ART) owing to a reservoir of latently infected cells that contain replication-competent virus<sup>1–4</sup>. Here, to better understand the mechanisms responsible for latency persistence and reversal, we used the interleukin-15 superagonist N-803 in conjunction with the depletion of CD8<sup>+</sup> lymphocytes in ART-treated macaques infected with simian immunodeficiency virus (SIV). Although N-803 alone did not reactivate virus production, its administration after the depletion of CD8<sup>+</sup> lymphocytes in conjunction with ART treatment induced robust and persistent reactivation of the virus in vivo. We found viraemia of more than 60 copies per ml in all macaques ( $n = 14$ ; 100%) and in 41 out of a total of 56 samples (73.2%) that were collected each week after N-803 administration. Notably, concordant results were obtained in ART-treated HIV-infected humanized mice. In addition, we observed that co-culture with CD8<sup>+</sup> T cells blocked the in vitro latency-reversing effect of N-803 on primary human CD4<sup>+</sup> T cells that were latently infected with HIV. These results advance our understanding of the mechanisms responsible for latency reversal and lentivirus reactivation during ART-suppressed infection.

HIV remains a major global health problem, leading to approximately 1.1 million deaths worldwide annually<sup>5</sup>. Despite the major declines in morbidity and mortality associated with the use of ART, there is neither a vaccine nor a cure for HIV infection. The inability to eradicate HIV infection with the current therapies is due to the presence of latently infected cells that contain integrated replication-competent virus that persist indefinitely in HIV-infected individuals undergoing ART and contribute to rebound viraemia when therapy is discontinued (that is, the viral reservoir)<sup>1–4</sup>. A key paradigm in the field of HIV cure—referred to as ‘shock and kill’<sup>6,7</sup>—supposes that the induction of virus expression (that is, ‘virus reactivation’) in these latently infected cells (‘shock’) followed by immune-mediated clearing (‘kill’) may substantially reduce the reservoir size and possibly lead to a functional cure for HIV infection. Unfortunately, no latency-reversing agent (LRA) tested to date has successfully perturbed the viral reservoir in clinical trials in humans. In particular, histone deacetylase inhibitors have not induced either robust virus reactivation or reduction of the viral reservoir in ART-treated HIV-infected individuals<sup>8–12</sup>. More encouragingly,

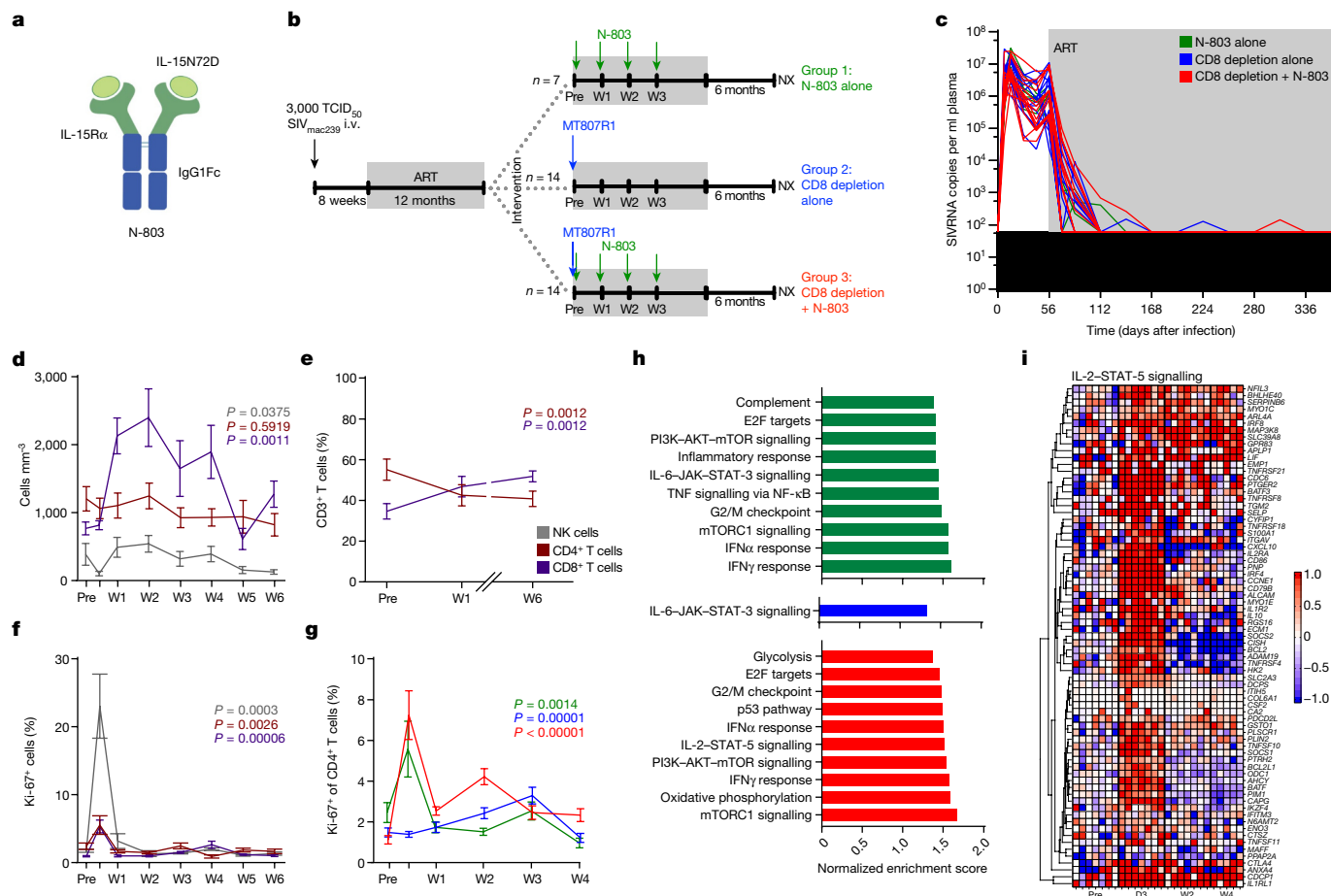
in SIV-infected ART-treated rhesus macaques (*Macaca mulatta*), treatment with toll-like receptor 7 (TLR7) agonists was linked to transient blips of plasma viraemia<sup>13</sup>. However, this result was not reproduced in further studies<sup>14,15</sup>. More recently, persistent remission was observed in a subset of ART-treated simian–human chimeric immunodeficiency virus (SHIV)-infected macaques that received the TLR7 agonist GS-9620 in combination with the broadly neutralizing PGT121 antibody<sup>16</sup>. In all, these published data indicate that novel, more potent approaches for latency reversal are needed to achieve a functional cure for HIV infection.

## SIV and SHIV infection of rhesus macaques

Infection of rhesus macaques with SIV or SHIV is the most widely used animal model for the study of the mechanisms by which the viral reservoir is established and maintained under ART and to preclinically test interventions aimed at reducing the viral reservoir in vivo<sup>17</sup>. A previous study demonstrated that the depletion of CD8<sup>+</sup> lymphocytes

<sup>1</sup>Emory Vaccine Center and Yerkes National Primate Research Center, Emory University, Atlanta, GA, USA. <sup>2</sup>Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA.

<sup>3</sup>Vaccine and Gene Therapy Institute, Oregon Health & Science University, Beaverton, OR, USA. <sup>4</sup>International Center for the Advancement of Translational Science, Division of Infectious Diseases, Center for AIDS Research, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>5</sup>University of North Carolina HIV Cure Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>6</sup>Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA. <sup>7</sup>AIDS and Cancer Virus Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. <sup>8</sup>NantKwest, Culver City, CA, USA. <sup>9</sup>Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, OR, USA. <sup>10</sup>Emory + Children's Center for Childhood Infections and Vaccines, Atlanta, GA, USA. <sup>11</sup>These authors jointly supervised this work: Ann Chahroudi, Guido Silvestri. \*e-mail: gsilves@emory.edu



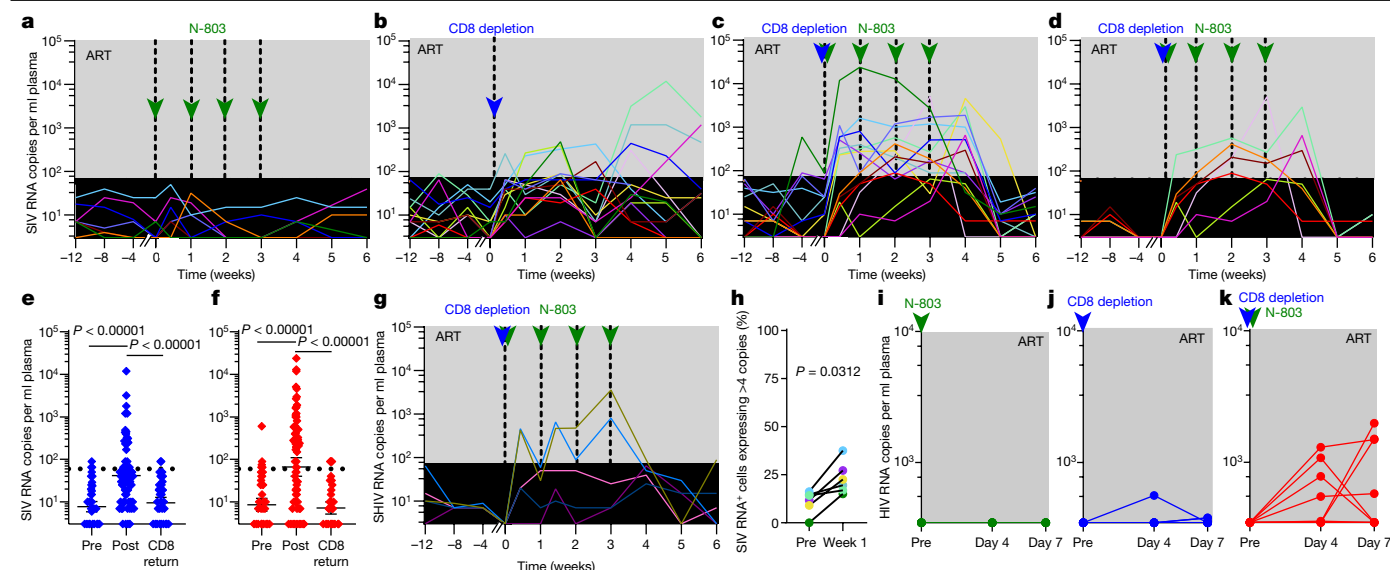
**Fig. 1 | Study design and phenotypic and transcriptomic effects of N-803 with or without CD8 depletion in rhesus macaques.** **a**, IL-15 superagonist N-803 structure. i.v., intravenous; TCID<sub>50</sub>, 50% tissue-culture infectious dose. **b**, Study design. Arrows indicate the administration of 100 μg kg<sup>-1</sup> N-803 (green) and 50 mg kg<sup>-1</sup> MT807R1 (blue). **c**, Plasma viral load pre-intervention ( $n = 35$  macaques), including infection and initiation of ART (grey bar). Limit of detection is 60 copies of SIV RNA per ml of plasma (black bar). **d**, Mean peripheral CD4<sup>+</sup> T cell (maroon), CD8<sup>+</sup> T cell (purple) and NK cell (grey) count in the lymph node. **e**, Percentage of CD4<sup>+</sup> and CD8<sup>+</sup> T cells in the lymph node. **f**, Ki-67 expression in cellular subsets after intervention with N-803 ( $n = 7$  macaques). **g**, Ki-67 expression in bulk CD4<sup>+</sup> T cells after treatment with only N-803 (green,  $n = 7$  macaques), CD8 depletion alone (blue,

$n = 14$  macaques), and CD8 depletion combined with the administration of N-803 (red,  $n = 14$  macaques). **h**, Gene-set enrichment analysis of RNA-sequencing data from bulk CD4<sup>+</sup> T cells comparing gene sets enriched on day 3 after intervention with N-803 alone (green,  $n = 7$  macaques), CD8 depletion alone (blue,  $n = 7$  macaques) or CD8 depletion combined with N-803 treatment (red,  $n = 7$  macaques). **i**, Heat map of the enriched genes in bulk CD8<sup>+</sup> T cells in the IL-2 and STAT-5 signalling gene set after administration of N-803 alone ( $n = 7$  macaques). Data are mean  $\pm$  s.e.m. Two-sided Kruskal–Wallis tests (**d**, **f**) and Friedman tests (**e**, **g**) were used to compare values after the interventions to the pre-intervention baseline and approximate  $P$ -value summaries are provided.

in SIV-infected ART-treated macaques was consistently followed by increased plasma viraemia, thus indicating that these cells contribute to viral suppression under ART<sup>18</sup>. Although the precise mechanisms responsible for this observation remain unclear, phylogenetic analysis of the rebounding virus suggested that silencing of virus transcription contributes to this antiviral effect. On the basis of these observations, we hypothesized that the depletion of CD8<sup>+</sup> lymphocytes combined with LRAs may enhance virus production under ART. As shown in Fig. 1a, the IL-15 superagonist N-803 is a complex of a mutant IL-15 and a dimeric IL-15 receptor αSu/Fc fusion protein<sup>19</sup>. The engineered structure is at least 25 times more biologically potent than IL-15 as it mimics transpresentation, and the IgG–Fc component confers improved in vivo safety and bioavailability<sup>20,21</sup>. In the setting of ART-suppressed lentiviral infection, N-803 may target the residual virus pool through its ability to act in vitro as a potent LRA and to strengthen the antiviral immune responses mediated by T and natural killer (NK) cells<sup>22</sup>.

The current study included a total of 35 SIV-infected macaques that started ART at day 56 after infection and were treated for at least 1 year before any further intervention. The macaques were divided in three groups as follows (Fig. 1b): 7 macaques were treated with 4 weekly doses

of 100 μg kg<sup>-1</sup> of N-803 (group 1, N-803 alone), 14 macaques received 1 dose of the CD8-depleting antibody, MT807R1 (anti-CD8α) at 50 mg kg<sup>-1</sup> intravenously (group 2, CD8 depletion alone) and 14 macaques received 4 weekly doses of N-803 starting at the time of CD8 depletion (group 3, CD8 depletion with N-803). After reconstitution of CD8<sup>+</sup> T cells (defined as more than 100 CD8<sup>+</sup> T cells per μl of blood), 7 macaques in groups 2 and 3 received four additional weekly administrations of N-803. Peripheral blood samples, lymph node and rectal biopsies were collected at various time points before, during and after these interventions. All macaques underwent analytical treatment interruption at week 3 after either CD8<sup>+</sup> T cell reconstitution or the last N-803 treatment. As shown in Fig. 1c, all macaques showed suppression of viraemia after 1 year of ART, with plasma viral loads below the detectable limit of our standard assay (60 copies per ml of plasma)<sup>23</sup> at the time of the additional interventions in 33 out of 35 macaques (94.3%). We also measured residual plasma viraemia using an ultrasensitive assay (limit of detection of 3 copies per ml)<sup>24–26</sup> at three monthly sampling points before the interventions. Viraemia was below 3 copies per ml in 19 out of 35 ART-treated macaques (52.3%), and 26 out of 35 macaques (74.3%) showed levels of residual viraemia of  $\leq 10$  copies per ml at the time of



**Fig. 2 | SIV and HIV reactivation after CD8 depletion combined with N-803 treatment.** **a–c**, Plasma viral loads after intervention with N-803 alone ( $n = 7$  macaques) (**a**), CD8 depletion alone ( $n = 14$  macaques) (**b**) or CD8 depletion combined with N-803 ( $n = 14$  macaques) (**c**). **d**, Longitudinal plasma viral loads for macaques with fully suppressed viral load ( $< 3$  copies per ml of plasma) before CD8 depletion combined with administration of N-803. **e, f**, Comparison of viral load pre-intervention (pre), post-intervention when CD8<sup>+</sup> T cells are  $< 100$  cells per  $\mu$ l blood (post) and during CD8<sup>+</sup> T cell reconstitution  $> 100$  cells per  $\mu$ l blood (CD8 return) in macaques that underwent CD8 depletion alone ( $n = 14$  macaques) (**e**) or CD8 depletion with N-803 ( $n = 14$  macaques) (**f**). Data are mean  $\pm$  s.e.m. **g**, Plasma viral loads after CD8 depletion with N-803 administration in SHIV<sub>SFI62P3</sub>-infected macaques

after 6 months of ART ( $n = 5$  macaques). Viral suppression of  $< 60$  copies per ml is shown as a black bar (**a–d, g**) or dashed line (**e, f**) and the limit of detection for these assays was 3 copies per ml of plasma. **h**, RNAscope determination of the percentage of SIV RNA<sup>+</sup> lymph node cells expressing high levels ( $> 4$  copies) of viral RNA per cell 1 week after CD8 depletion combined with N-803 treatment. **i–k**, Plasma viral loads of HIV-infected, ART-treated humanized mice treated with N-803 alone (green,  $n = 7$  mice) (**i**), CD8 depletion alone (blue,  $n = 8$  mice) (**j**) or CD8 depletion with N-803 administration (red,  $n = 8$  mice) (**k**). The limit of detection was 346 copies per ml. Statistical significance was calculated using a two-sided Kruskal–Wallis test (**e, f**) or Wilcoxon signed-rank test (**h**). A key for the macaque ID codes is provided in Extended Data Fig. 5.

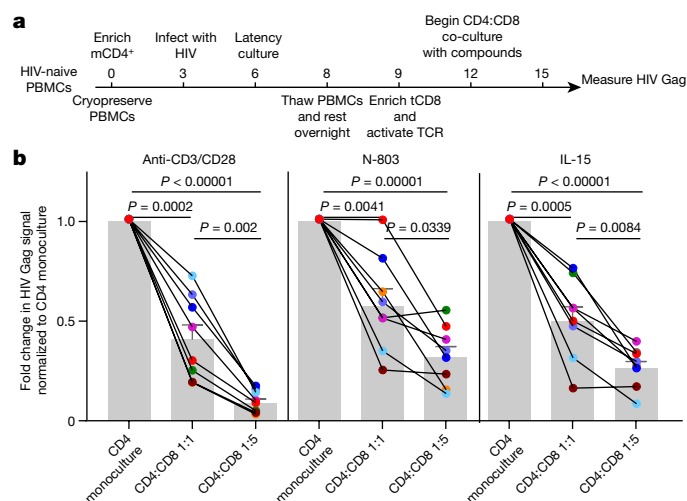
intervention (Extended Data Table 1). These results indicate that the level of virus suppression observed in our cohort of macaques was in most cases comparable to that of long-term ART-treated HIV-infected individuals<sup>27–29</sup>.

As shown in Extended Data Fig. 1a–d and consistent with previous studies<sup>18,30</sup>, treatment with anti-CD8 $\alpha$  MT807R1, with or without N-803, depleted on average 99.1% of CD3<sup>+</sup>CD8<sup>+</sup> T cells in peripheral blood, 97.9% in lymph nodes and 99.5% in rectal biopsies. In addition, treatment with MT807R1 alone depleted 93.2% of NK cells in peripheral blood (Extended Data Fig. 1e). As expected on the basis of previous studies, N-803 administration alone resulted in the expansion of CD8<sup>+</sup> T cells in the blood and lymph nodes (Fig. 1d, e), as well as increased proliferation of peripheral CD8<sup>+</sup> T cells, CD4<sup>+</sup> T cells and NK cells (Fig. 1f). Of note, although CD8 depletion alone did not result in a rapid increase in CD4<sup>+</sup> T cell proliferation (as measured by Ki-67 expression), the combination of CD8 depletion and treatment with N-803 led to a significant increase in CD4<sup>+</sup> T cell proliferation (Fig. 1g and Extended Data Fig. 2r–v). The frequency of CD4<sup>+</sup> T cells co-expressing Ki-67 and the HIV and SIV coreceptor CCR5—potential target cells for infection—was significantly increased across all groups by the third week after intervention (Extended Data Fig. 2l). Additionally, CD8 depletion with or without N-803 administration resulted in the expansion of effector memory CD4<sup>+</sup> T cells and increased PD-1 expression on CD4<sup>+</sup> T cells (Extended Data Fig. 2f, g).

To better characterize the biological effects of N-803, we conducted a transcriptional analysis using RNA sequencing of sorted CD4<sup>+</sup> T cells collected before intervention and day 3, week 2 and week 4 after the start of intervention. Regardless of concurrent CD8 depletion, N-803 induced a significant upregulation of gene sets associated with cell cycling and proliferation, activation, antiviral responses and cell signalling (Fig. 1h). In CD8<sup>+</sup> T cells, N-803 resulted in significant enrichment of genes in the IL-2 and STAT-5 signalling gene set, which is also indicative

of IL-15 signalling, as the receptor for this cytokine shares two out of three subunits with IL-2 and uses STAT-5 as the key adaptor molecule (Fig. 1i). In addition, we examined the expression of 25 genes specifically involved in the host–virus interaction during SIV infection and found that N-803 administration induced a consistent and transient upregulation of APOBEC3 in CD4<sup>+</sup> T cells, CD8<sup>+</sup> T cells and NK cells (Extended Data Fig. 3a–c).

As shown in Fig. 2a and Extended Data Table 1, administration of N-803 was not associated with an increase of plasma viraemia of  $> 60$  copies per ml in any of the treated macaques, indicating that the IL-15 superagonist is not sufficient to exert an *in vivo* LRA effect in ART-treated SIV-infected macaques when used alone. As expected on the basis of previous studies<sup>18</sup>, macaques undergoing CD8 depletion alone showed a moderate but significant increase in virus production, with plasma viraemia of  $> 60$  copies per ml detected in 11 out of 14 macaques (78.6%) and 18 out of 56 samples (32.1%) collected weekly after CD8 depletion (Fig. 2b). Viraemia of  $> 1,000$  copies per ml was observed in 2 out of 14 macaques (14.2%) and 2 out of 56 (3.6%) of the same samples (Fig. 2b). In all cases, the level of virus production returned to below 60 copies per ml of plasma at the time of CD8<sup>+</sup> T cell reconstitution (Fig. 2e). Overall, the level of increased viraemia observed in this study was consistent with previous studies<sup>18</sup>, even though the magnitude of virus production after CD8 depletion was slightly less marked, possibly related to the longer period of ART treatment (12 months compared with 2–8 months)<sup>18</sup>. Notably, macaques treated with N-803 during CD8 depletion showed highly robust and persistent levels of virus production, with viraemia of  $> 60$  copies per ml detected in 14 out of 14 macaques (100%) and 41 out of 56 samples (73.2%) and viraemia of  $> 1,000$  copies per ml observed in 6 out of 14 macaques (42.9%) and 13 out of 56 samples (23.2%) (Fig. 2c). We emphasize that all seven macaques with full suppression of virus production at the time of intervention with CD8 depletion and N-803 administration (that is,



**Fig. 3 | In vitro co-culture of latently infected human CD4<sup>+</sup> T cells with autologous CD8<sup>+</sup> T cells results in decreased expression of HIV Gag during LRA administration.** **a**, Schematic of HIV latency model. Memory CD4<sup>+</sup> T cells (mCD4<sup>+</sup>) were enriched on day 0, infected in vitro on day 3 with HIV<sub>89.6</sub> and maintained in saquinavir. On day 6, TGFβ, IL-7, conditioned medium from the H-80 feeder cell line and additional antiretroviral drugs were added to the culture. Cryopreserved autologous peripheral-blood mononuclear cells (PBMCs) were thawed on day 8 and rested overnight before enriching for total CD8<sup>+</sup> (tCD8<sup>+</sup>) cells and then TCR-activated for 3 days. On day 12, HIV-infected mCD4<sup>+</sup> and TCR-activated total CD8<sup>+</sup> cells were co-cultured in a 1:1 or 1:5 ratio in the presence of anti-CD3/CD28 antibody, N-803 or recombinant IL-15 until day 15. **b**, HIV Gag<sup>+</sup> CD4<sup>+</sup> T cell frequency was quantified using flow cytometry and the frequency in co-cultures was calculated as a fold change compared to CD4 T cell monoculture. Each colour represents a unique donor ( $n = 8$  biologically independent samples) and data are mean  $\pm$  s.e.m. (indicated by the grey bars). Statistical significance was measured using a matched one-way analysis of variance (ANOVA).

repeated viral load measurements <3 copies per ml of plasma) demonstrated clear virus reactivation with detectable levels in 26 out of 28 time points 1 week after each N-803 administration, and >60 copies per ml in 16 out of 28 of the same samples (Fig. 2d). Similar results were obtained in a smaller pilot study in which N-803 administration during CD8 depletion was performed in five ART-suppressed SHIV<sub>SF162P3</sub>-infected macaques (Fig. 2g). The level of viraemia observed in SIV-infected macaques treated with combined CD8 depletion and N-803 administration during long-term ART is higher and more persistent than the results of previous shock-and-kill cure strategies tested in humans and nonhuman primates<sup>8–13,31</sup>.

After the last treatment with N-803, the level of viraemia rapidly decreased coincident with the reconstitution of CD8<sup>+</sup> T cells, and all macaques returned to <60 copies per ml by week 6 after CD8 depletion and N-803 administration (Fig. 2c, f). As expected, CD8<sup>+</sup> T cell reconstitution was faster in CD8-depleted macaques co-treated with N-803 (Extended Data Fig. 5e–g), as the IL-15 superagonist enhances CD8<sup>+</sup> T cell proliferation (Fig. 1d–f).

We next investigated the correlates of virus reactivation in ART-treated SIV-infected macaques that underwent CD8 depletion with N-803, and observed that the post-depletion viral load (day 3 to week 6) was negatively correlated with the frequency of peripheral CD8<sup>+</sup> T cells (Extended Data Fig. 5b). Additionally, the area under the curve of virus production during CD8 depletion and administration of N-803 was directly correlated with pre-intervention viraemia (Extended Data Fig. 5d). Of note, no correlation was found between the level of virus production after CD8 depletion and/or N-803 treatment and either the size of the peripheral blood DNA reservoir (measured as the fraction of SIV DNA<sup>+</sup> CD4<sup>+</sup> T cells) or the level of CD4<sup>+</sup> T cell activation (measured as Ki-67 or PD-1 expression on CD4<sup>+</sup> T cells) (data not shown). To

assess whether combined CD8 depletion with N-803 administration induced SIV production in lymphoid tissues, we next analysed the levels of SIV RNA in the lymph nodes using the RNAscope technology at pre-intervention and day 7 after intervention in 5 representative macaques that were both depleted of CD8<sup>+</sup> lymphocytes and treated with N-803. As shown in Fig. 2h, and consistent with the measurements of plasma viraemia, we found a statistically significant increase in the percentage of SIV RNA<sup>+</sup> cells with high levels of SIV RNA after intervention. No changes were observed in the level of SIV RNA in peripheral CD4<sup>+</sup> T cells (Extended Data Fig. 4a–c), suggesting that lymphoid tissues are the main source of reactivated virus after the combined CD8 depletion and administration of N-803.

## HIV infection of humanized mice

To confirm the virus reactivation induced by the combined depletion of CD8<sup>+</sup> lymphocytes and administration of N-803 in SIV-infected ART-treated macaques in an in vivo model using HIV, we next conducted a similarly designed experiment using bone-marrow–liver–thymus (BLT) humanized mice infected with HIV-1<sub>JR-CSF</sub> and treated with ART. As shown in Fig. 2i–k, HIV-infected humanized mice showed markedly similar results to those obtained in SIV-infected macaques, with no plasma virus reactivation after the administration of N-803 alone, a moderate level of virus reactivation after CD8 depletion alone, and a robust level of virus reactivation in 7 out of 8 (87.5%) humanized mice that were depleted of CD8<sup>+</sup> lymphocytes and treated with N-803. Furthermore, we also noted a statistically significant increase in the levels of cell-associated HIV RNA in the spleen and human-derived thymus of humanized mice that were depleted of CD8<sup>+</sup> lymphocytes and treated with N-803 (Extended Data Fig. 6b).

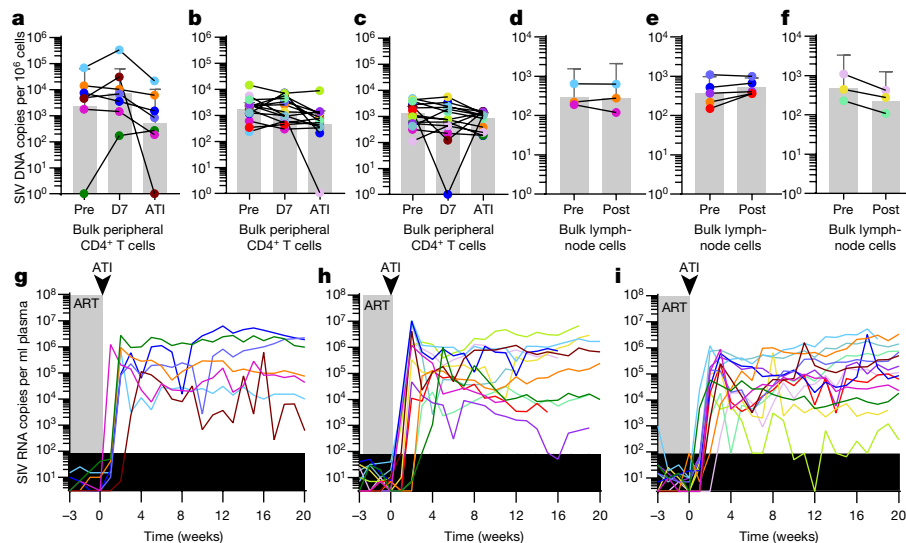
## CD8<sup>+</sup> T cells inhibit latency reversal in vitro

The combined data obtained in SIV-infected macaques and HIV-infected humanized mice indicate that the strong virus reactivation activity attributed to N-803 is revealed only in the absence of CD8<sup>+</sup> T cells, thus delineating a novel mechanism of latency maintenance and/or inhibition of latency reversal mediated by CD8<sup>+</sup> lymphocytes. To recapitulate this observation in a reductionist in vitro model of HIV latency in human cells, we used the recently developed latency and reversal assay (LARA)<sup>32</sup> to evaluate how CD8<sup>+</sup> T cells affect the virus reactivation activity of N-803 in autologous memory CD4<sup>+</sup> T cells latently infected with HIV<sub>89.6</sub> (Fig. 3a). Of note, this assay was conducted using cells derived from HIV-negative donors, thus HIV-specific cytotoxic CD8<sup>+</sup> T cells were absent. Whereas N-803 (and its biological counterpart, IL-15) reactivates HIV expression in latently infected CD4<sup>+</sup> T cell monocultures, co-culture with activated CD8<sup>+</sup> T cells significantly suppresses this ability (Fig. 3b). These data indicate that CD8<sup>+</sup> T cells effectively suppress the latency-reversing activity of N-803, and therefore confirm the discovery of a previously unrecognized CD8<sup>+</sup> T-cell-mediated activity that contributes to the maintenance of lentivirus latency in vivo in primates.

## Virus sequence analysis

To investigate the viral dynamics associated with reactivation after CD8 depletion and administration of N-803 in SIV-infected ART-treated macaques, we performed a longitudinal sequence analysis of plasma virus using single-genome PCR amplification of the SIV<sub>mac239</sub>-derived *env* genes. The viral *env* was sequenced at three pivotal time points: day 7 after infection (at peak viraemia), day 56 after infection—that is, immediately before ART initiation (pre-ART)—and during peak virus reactivation. We conducted this analysis in six macaques that showed robust virus reactivation (plasma viraemia of >800 copies per ml) (Fig. 2c). Extended Data Figure 7 shows phylogenetic analysis of the translated Env amino acid sequences of the circulating viruses.





**Fig. 4 | CD8 depletion combined with N-803 administration does not decrease the size of the latent SIV viral reservoir.** **a–f**, Copies of cell-associated SIV DNA per  $10^6$  cells was determined in peripheral CD4<sup>+</sup> T cells (**a–c**) and frozen bulk lymph node cells (**d–f**) in macaques treated with N-803 alone ( $n = 7$  (**a**) and  $n = 3$  (**d**) macaques), CD8 depletion alone ( $n = 14$  (**b**) and  $n = 5$  (**e**) macaques), and CD8 depletion with N-803 treatment ( $n = 14$  (**c**) and  $n = 3$  (**f**) macaques). Two-sided Friedman tests (**a–c**) and Wilcoxon signed-rank tests (**d–f**) were used to determine statistical significance between pre-treatment

and post-treatment time points. For all comparisons,  $P > 0.05$ . Sample mean  $\pm$  s.e.m. are indicated by grey bars. **g–i**, Viral rebound after interruption of ART (indicated by arrowheads) in macaques that received N-803 alone ( $n = 7$  macaques) (**g**), CD8 depletion alone ( $n = 13$  macaques) (**h**) or CD8 depletion with N-803 ( $n = 14$  macaques) (**i**). ATI, analytical treatment interruption. The limit of detection was  $< 3$  copies of SIV RNA per ml plasma. The black bar indicates viral loads  $< 60$  copies per ml and the grey box indicates ART.

The peak viral load Env sequences were homogeneous and nearly identical to the input SIV<sub>mac239</sub> sequence; however, the diversity and number of informative sites at subsequent time points were limited, such that sequences could not be clustered on the basis of time point with significant bootstrap support. The diversity at each time point, although limited, was quantified by determining the number of amino acid differences from the input SIV<sub>mac239</sub> sequence. Extended Data Figure 8a shows that, for all macaques, the peak viraemia Env sequences are the least different from the input virus, as expected, and increased divergence was observed at the pre-ART time point and after reactivation. We next calculated the average number of amino acid differences from SIV<sub>mac239</sub> at each time point, and compared these with contemporaneous plasma viraemia for each macaque in a correlation matrix. The only significant association was a direct relationship between Env divergence and plasma viraemia during reactivation (Extended Data Fig. 8b). Finally, Extended Data Fig. 9 shows the location of sequence changes at each time point using highlighter plots of the Env amino acid sequences. Overall, this analysis supports the hypothesis that CD8 depletion and administration of N-803 induces robust reactivation of a diverse population of viral variants. As no signs of virus evolution emerged from the longitudinal sequence analysis during high levels of latency reversal, it is unlikely that the rebounding virus is a product of de novo viral replication. Supporting this hypothesis, the combined CD8 depletion and administration of N-803 did not increase the levels of two-long-terminal repeat circles—which are considered a marker of recent lentivirus infection—in peripheral-blood mononuclear cells (data not shown).

### Analytic treatment interruption

To determine whether the interventions used induced a decrease in the virus reservoir, we first longitudinally measured the level of cell-associated SIV DNA in blood and lymph nodes. As shown in Fig. 4a–f, none of the experimental groups showed significant changes in either the total fraction of circulating CD4<sup>+</sup> T cells or the calculated fraction of lymph-node-derived cells that contained SIV DNA. To functionally assess the effect of the treatment regimens on the reservoir size, we

performed an analytical treatment interruption in all macaques three weeks after either CD8 reconstitution and/or the last N-803 treatment. As shown in Fig. 4g–i, all macaques rebounded within three weeks of ART interruption and most macaques sustained high viral loads until the time of necropsy. It should be noted that in the current study, ART was initiated at day 56 after infection, thus substantially later than in other published macaque studies that included analytical treatment interruption, and thus in the setting of a larger and more-disseminated reservoir<sup>16,33</sup>. The rapid rebound after ART interruption was therefore not unexpected as the experimental design was focused on assessing the shock effect of CD8 depletion combined with N-803 administration, with no anticipated effect on the reservoir size in the absence of an intervention aimed at clearing the cells that have reactivated SIV production (kill phase of the shock-and-kill approach). The absence of a decrease in the level of SIV DNA<sup>+</sup> cells after the combined CD8 depletion and administration of N-803 may be due to the lack of CD8<sup>+</sup> T-cell-mediated clearance of cells that have reactivated virus expression and/or the N-803-mediated proliferative expansion of infected CD4<sup>+</sup> T cells that have survived the events of virus reactivation.

### Discussion

The current paradigm for shock-and-kill interventions for an HIV cure suggests that reactivation of virus transcription in latently infected cells is the first essential step to eliminate the persistent reservoir of replication-competent virus in ART-treated HIV-infected individuals. In this study, we have shown that the administration of the IL-15 superagonist N-803 in both SIV-infected macaques and HIV-infected humanized mice induces a highly robust and persistent reversal of latency only in the setting of CD8<sup>+</sup> lymphocyte depletion, thus suggesting a substantial role for CD8<sup>+</sup> lymphocytes in suppressing the LRA effect of N-803. Notably, this previously undescribed role of CD8<sup>+</sup> lymphocytes in maintaining virus latency was fully reproduced in an in vitro experimental approach that involved the co-culture of activated, unprimed CD8<sup>+</sup> T cells with autologous, latently HIV-infected human primary CD4<sup>+</sup> T cells. Here we use a novel approach to manipulate latently infected cells and have independently confirmed this method in the two best-validated and



most widely used in vivo models for HIV cure interventions and then recapitulated the results in an in vitro experimental system of virus latency. In addition to this conceptual advance, this study defines a robust shock approach that could provide a key experimental system to directly compare and contrast the efficacy of different kill interventions in vivo, including those that may act in a CD8<sup>+</sup> T-cell-independent manner (such as neutralizing antibodies, CD4 mimetics or immunotoxins). Further studies that aim to identify the specific molecular pathways used by CD8<sup>+</sup> T cells to promote latency may allow the suppression of this activity, and therefore enable the full use of the virus-reactivating potential of N-803 or other LRAs in the clinical setting without depleting CD8<sup>+</sup> lymphocytes.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1946-0>.

- Chun, T.-W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
- Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
- Wong, J. K. et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
- Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4<sup>+</sup> T cells. *Nat. Med.* **9**, 727–728 (2003).
- UNAIDS. *UNAIDS Data 2017*. (2017).
- The International AIDS Society Scientific Working Group on HIV Cure. Towards an HIV cure: a global scientific strategy. *Nat. Rev. Immunol.* **12**, 607–614 (2012).
- Archin, N. M. et al. Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487**, 482–485 (2012).
- Archin, N. M. et al. HIV-1 expression within resting CD4<sup>+</sup> T cells after multiple doses of vorinostat. *J. Infect. Dis.* **210**, 728–735 (2014).
- Rasmussen, T. A. et al. Panobinostat, a histone deacetylase inhibitor, for latent-virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: a phase 1/2, single group, clinical trial. *Lancet HIV* **1**, e13–e21 (2014).
- Søgaard, O. S. et al. The deipeptide romidepsin reverses HIV-1 latency in vivo. *PLoS Pathog.* **11**, e1005142 (2015).
- Elliot, J. H. et al. Activation of HIV transcription with short-course vorinostat in HIV-infected patients on suppressive antiretroviral therapy. *PLoS Pathog.* **10**, e1004473 (2014).
- Elliot, J. H. et al. Short-term administration of disulfiram for reversal of latent HIV infection: a phase 2 dose-escalation study. *Lancet HIV* **2**, e520–e529 (2015).
- Lim, S.-Y. et al. TLR7 agonists induce transient viremia and reduce the viral reservoir in SIV-infected rhesus macaques on antiretroviral therapy. *Sci. Transl. Med.* **10**, eaa04521 (2018).
- Del Prete, G. Q. et al. TLR7 agonist administration to SIV-infected macaques receiving early initiated cART does not induce plasma viremia. *JCI Insight* **4**, e127717 (2019).
- Bekerman, E. et al. PD-1 blockade and TLR7 activation lack therapeutic benefit in chronic simian immunodeficiency virus-infected macaques on antiretroviral therapy. *Antimicrob. Agents Chemother.* **63**, e01163-19 (2019).
- Borducchi, E. N. et al. Antibody and TLR7 agonist delay viral rebound in SHIV-infected monkeys. *Nature* **563**, 360–364 (2018).
- Nixon, C. C., Mavigner, M., Silvestri, G. & Garcia, J. V. In vivo models of human immunodeficiency virus persistence and cure strategies. *J. Infect. Dis.* **215**, S142–S151 (2017).
- Cartwright, E. K. et al. CD8<sup>+</sup> lymphocytes are required for maintaining viral suppression in SIV-infected macaques treated with short-term antiretroviral therapy. *Immunity* **45**, 656–668 (2016).
- Xu, W. et al. Efficacy and mechanism-of-action of a novel superagonist interleukin-15: interleukin-15 receptor  $\alpha$ Su/Fc fusion complex in syngeneic murine models of multiple myeloma. *Cancer Res.* **73**, 3075–3086 (2013).
- Han, K. P. et al. IL-15:IL-15 receptor  $\alpha$  superagonist complex: high-level co-expression in recombinant mammalian cells, purification and characterization. *Cytokine* **56**, 804–810 (2011).
- Rhode, P. R. et al. Comparison of the superagonist complex, ALT-803, to IL15 as cancer immunotherapeutics in animal models. *Cancer Immunol. Res.* **4**, 49–60 (2016).
- Jones, R. B. et al. A subset of latency-reversing agents expose HIV-infected resting CD4<sup>+</sup> T-cells to recognition by cytotoxic T-lymphocytes. *PLoS Pathog.* **12**, e1005545 (2016).
- Hofmann-Lehmann, R. et al. Sensitive and robust one-tube real-time reverse transcriptase-polymerase chain reaction to quantify SIV RNA load: comparison of one-versus two-enzyme systems. *AIDS Res. Hum. Retroviruses* **16**, 1247–1257 (2010).
- Del Prete, G. Q. et al. Effect of suberoylanilide hydroxamic acid (SAHA) administration on the residual virus pool in a model of combination antiretroviral therapy-mediated suppression in SIVmac239-infected Indian rhesus macaques. *Antimicrob. Agents Chemother.* **58**, 6790–6806 (2014).
- Hansen, S. G. et al. Immune clearance of highly pathogenic SIV infection. *Nature* **502**, 100–104 (2013).
- Li, H. et al. Envelope residue 375 substitutions in simian–human immunodeficiency viruses enhance CD4 binding and replication in rhesus macaques. *Proc. Natl Acad. Sci. USA* **113**, E3413–E3422 (2016).
- Dornadula, G. et al. Residual HIV-1 RNA in blood plasma of patients taking suppressive highly active antiretroviral therapy. *J. Am. Med. Assoc.* **282**, 1627–1632 (1999).
- Maldarelli, F. et al. ART suppresses plasma HIV-1 RNA to a stable set point predicted by pretherapy viremia. *PLoS Pathog.* **3**, e46 (2007).
- Chun, T. W. et al. Relationship between residual plasma viremia and the size of HIV proviral DNA reservoirs in infected individuals receiving effective antiretroviral therapy. *J. Infect. Dis.* **204**, 135–138 (2011).
- Chowdhury, A. et al. Differential impact of in vivo CD8<sup>+</sup> T lymphocyte depletion in controller versus progressor simian immunodeficiency virus-infected macaques. *J. Virol.* **89**, 8677–8686 (2015).
- Spivak, A. M. et al. A pilot study assessing the safety and latency-reversing activity of disulfiram in HIV-1-infected adults on antiretroviral therapy. *Clin. Infect. Dis.* **58**, 883–890 (2014).
- Kulpa, D. A. et al. Differentiation to an effector memory phenotype potentiates HIV-1 latency reversal in CD4<sup>+</sup> T cells. *J. Virol.* **93**, e00969-19 (2019).
- Okoye, A. A. et al. Early antiretroviral therapy limits SIV reservoir establishment to delay or prevent post-treatment viral rebound. *Nat. Med.* **24**, 1430–1440 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

# Article

## Methods

### Rhesus macaques, SIV-infection, ART, CD8 depletion and administration of N-803

This study was conducted using a cohort of 35 Indian-origin rhesus macaques housed at Yerkes National Primate Research Center (male and female, 2–3 years of age at the start of the study). All macaques were Mamu\*B07<sup>+</sup> and Mamu\*B17<sup>+</sup>; the following macaques were Mamu\*A01<sup>+</sup>: 77\_13, RFr15, 208\_13, RPb16, RJt15, RHv15, RAu15, RNa16, RNz15, ROr15, RRB16, RSt15, RAK16, RU515, RYe16, RVz15, RBn16 and RRn16. All procedures were approved by the Emory University Institutional Animal Care and Use Committee (IACUC) and animal care facilities are accredited by the US Department of Agriculture and the Association for Assessment and Accreditation of Laboratory Animal Care International.

Rhesus macaques were infected intravenously with 3,000 TCID<sub>50</sub> of SIV<sub>mac239</sub> (*nef*open). SIV<sub>mac239</sub> stock was titrated in vitro for viral infectivity by standard end-point titration on CEMx174 cells. The TCID<sub>50</sub> was calculated using a previously published method<sup>34</sup>. All macaques were put on a three-drug ART regimen at 8 weeks after SIV infection. Tenofovir disoproxil fumarate (5.1 mg kg<sup>-1</sup> per day or tenofovir, 20 mg kg<sup>-1</sup> per day) and emtricitabine (40 mg kg<sup>-1</sup> per day) were both provided by Gilead Pharmaceuticals. Dolutegravir (2.5 mg kg<sup>-1</sup> per day) was provided by ViiV Pharmaceuticals. Drugs were administered daily by subcutaneous injection.

After over 12 months of ART, macaques were assigned to intervention groups. Age, weight, sex, A01 status, peak post-infection viral load and time to suppression after ART were all controlled for when allocating macaques to intervention groups. Group sizes were determined using a power analysis based on previous studies. No blinding was performed in this study. One dose of the anti-CD8 $\alpha$ -depleting antibody, MT807R1 (50 mg kg<sup>-1</sup>), was administered to 28 macaques. The initial 15 macaques that received the depleting antibody received the administration intravenously. Owing to safety concerns, the administration was changed to subcutaneous for the remaining 13 macaques. There was no observable effect of the different administration routes on the efficacy of depletion.

At the start of the intervention, 21 macaques received a dose of N-803 either in addition to CD8 depletion ( $n = 14$ ) or as a single treatment ( $n = 7$ ). N-803 was administered subcutaneously in a cycle of 100  $\mu$ g kg<sup>-1</sup> once a week for 4 consecutive weeks.

The study design included a subsequent four-dose administration of N-803 in seven macaques of groups 1 and 2 that was carried out at the time of CD8<sup>+</sup> T cell reconstitution to potentially accelerate the recovery of these cells and improve their antiviral cytotoxic potential. As expected, this second cycle of N-803 induced a faster recovery of CD8<sup>+</sup> T cells (data not shown) and was associated with an increase in T cell activation and proliferation that was similar to the increased activation and proliferation observed after the first N-803 cycle (data not shown). The late administration did not result in an increase in plasma viraemia.

Macaques were interrupted of ART 3 weeks after either the last dose of N-803 or 3 weeks after the reconstitution of CD8<sup>+</sup> T cells (whichever occurred first). Plasma viral loads were monitored for about 6 months until necropsy was performed.

### CD8 depletion in combination with N-803 treatment in SHIV-infected ART-treated macaques (pilot study)

An additional five Indian-origin rhesus macaques were included in this study as part of a follow-up pilot study using a SHIV model of infection. These macaques were also housed at the Yerkes National Primate Research Center and all procedures were approved by the Emory University IACUC. Macaques were infected intrarectally with high-dose SHIV<sub>SF162P3+</sub>, administered as a 1:50 dilution of a 2,032 TCID<sub>50</sub> per ml, 10<sup>9</sup> RNA copies per ml, 182 ng ml<sup>-1</sup> P27 stock. All macaques were placed on the same tenofovir, emtricitabine and dolutegravir ART

regimen 12 weeks after SHIV infection. After 6 months of ART, macaques received one dose of MT-807R1 at 50 mg kg<sup>-1</sup> subcutaneously. N-803 was administered subcutaneously in a cycle of 100  $\mu$ g kg<sup>-1</sup> once a week for 4 consecutive weeks starting at the time of CD8 depletion.

### Collection and processing of tissues from macaques

Blood, lymph nodes and rectal biopsies were collected longitudinally including at the time of necropsy and processed for further analyses as previously described<sup>18</sup>.

### Immunophenotypes of macaques by flow cytometry

Multiparametric flow cytometry was performed according to a standard protocol on PBMCs and lymph node mononuclear cells using fluorescently labelled monoclonal antibodies cross-reactive in rhesus macaques. The following antibodies were used at 37 °C for 30 min: CCR5 APC (BD Biosciences, 560748, clone 3A9) and CCR7 FITC (BD Biosciences, 561271, clone 150503), in addition to LIVE/DEAD aqua viability dye (ThermoFisher, L35957). The following antibodies were subsequently used at room temperature for 30 min: CD3 APC-Cy7 (BD Biosciences, 557757, clone SP34-2), CD4 BV650 (Biolegend, 317436, clone OKT4), CD8 $\alpha$  BV711 (Biolegend, 301044, clone RPA-T8), pure CD8 $\beta$  (Thermo, 14-5273-82, clone SIDI8BEE) conjugated to Pe-Cy5 via a kit (Innova Biosciences, 760-0010), CD45RA Pe-Cy7 (BD Biosciences, 561216, clone 5H9), CD62L PE (BD Biosciences, 341012, clone SK11), CD95 BV605 (Biolegend, 305628, clone DX2), PD-1 BV421 (Biolegend, 329920, clone EH12.2H7), CD16 BV421 (BD Biosciences, 562874, clone 3G8), CD20 PE-Cy5 (BD Biosciences, 555624, clone 2H7), CD14 PE-Cy7 (BD Biosciences, 557742, clone M5E2), NKG2A (also known as CD159) PE (Beckman Coulter, IM3291U, clone Z199), CD28 PE-Cy5.5 or ECD (Beckman Coulter, B24027 and 6607111, respectively, clone CD28.2), CD56 FITC (BD Biosciences, 340723, clone NCAM16.2). Additional panels included CD69 Pe-Cy5 (BD Biosciences, 555532, clone FN50) and CD25 APC (BD Biosciences, 555434, clone M-A251). Cells stained for Ki-67 were fixed and permeabilized with Perm II kit (BD Biosciences) before staining at room temperature for 30 min with Ki-67 AF700 (BD Biosciences, 561277, clone B56).

All flow cytometry specimens were acquired on an LSR II (BD Biosciences) equipped with fluorescence-activated cell sorting (FACS) software (FACS Diva), and analysis of the acquired data was performed using FlowJo software (TreeStar).

### Determination of plasma SIV RNA, and cell-associated RNA and DNA

For pre-intervention time points, quantitative PCR with reverse transcription (RT-qPCR) was performed to determine SIV plasma viral load as previously described<sup>23</sup> with a sensitivity of 60 copies per ml. For the three time points before intervention and all post-intervention time points, plasma SIV *gag* RNA levels were measured high-sensitivity assay formats<sup>24–26</sup>. Quantification of total cell-associated SIV<sub>mac239</sub> *gag* DNA was performed as previously described<sup>35</sup>. The number of *gag* DNA copies per 10<sup>6</sup> CD4<sup>+</sup> T cells was calculated by dividing the number of *gag* DNA copies per 10<sup>6</sup> PBMCs by the percentage of CD3<sup>+</sup>CD8<sup>+</sup>CD4<sup>+</sup> cells in the PBMC population. CD4<sup>+</sup> T cells were isolated from PBMCs using a CD4<sup>+</sup> T cell isolation kit (Miltenyi) and cell-associated RNA was measured as previously described<sup>36</sup>.

### In situ RNA analysis and quantification

Viral RNA (vRNA) detection using RNAscope and quantitative image analysis was performed on formaldehyde-fixed, paraffin-embedded tissue sections (5  $\mu$ m) as previously published<sup>37</sup>, with the following minor modifications: heat-induced epitope retrieval was performed by boiling slides in 1 $\times$  target retrieval (322000; ACD) for 30 min, followed by incubation at 40 °C with a 1:10 dilution of protease III (322337; ACD) in 1 $\times$  PBS for 20 min. Slides were incubated with the target probe SIV<sub>mac239</sub> (312811; ACD) for 2 h at 40 °C and amplification was performed

with RNAscope 2.5 HD Detection kits (322360; ACD) according to the manufacturer's instructions, with 0.5× wash buffer (310091; ACD) used between steps. The resulting signal was detected with Warp Red chromogen (WR806M; Biocare Medical). Slides were counterstained with CAT haematoxylin (CATHE-GL; Biocare Medical), mounted with Clearmount (17885-15; EMS) until dry, after which a coverslip was added and sealed using Permount (SP15-100; Fisher Scientific). Slides were scanned at 40× magnification on an Aperio AT2 (Leica Biosystems). RNAscope images were analysed for the total number of vRNA<sup>+</sup> cells per 10<sup>5</sup> total cells (quantitative) and the relative amount of vRNA present (semiquantitative) using the ISH module (v.2.2) within the Halo software (v.2.3.2089.27; Indica Labs). The relative amount of vRNA within a single cell was first estimated by quantifying the total area of the vRNA signal spot size (μm<sup>2</sup>). As the signal spot size is a function of several steps in the experimental procedures, module settings were established on concomitantly assayed, acutely infected SIV<sup>+</sup> control slides. To estimate the signal spot size of a single vRNA molecule, we measured the signal area (minimum, mean and maximum) of more than 10 identifiable individual virions within B cell follicles, which corresponds to two copies of vRNA, and multiplied this by 0.5. We set the vRNA minimum signal spot size within the analysis module to exclude detection of a single vRNA molecule and/or integrated viral DNA. Relative vRNA copy numbers that were present within vRNA<sup>+</sup> cells were calculated as (signal spot size within vRNA<sup>+</sup> cell (μm<sup>2</sup>))/(0.5 × mean signal size for a virion).

### FACS of live cells from macaques

Mononuclear cells isolated from blood were stained with LIVE/DEAD aqua viability dye (ThermoFisher, L35957), CD3 AF700 (BD Biosciences, 557917, clone SP34-2), CD4 BV650 (Biolegend, 317436, clone OKT4), CD8 APC-Cy7 (BD Biosciences, 557834, clone SK1), CD14 PB (Biolegend, 301828, clone M5E2), CD20 PB (Biolegend, 302328, clone 2H7) and CD16 BV421 (Biolegend, 302032, clone 3G8) for 30 min at room temperature. Aliquots of 50,000 CD4<sup>+</sup> T cells (live CD3<sup>+</sup>CD20<sup>−</sup>CD14<sup>−</sup>CD16<sup>−</sup>CD8<sup>−</sup>CD4<sup>+</sup>) and CD8<sup>+</sup> T cells (live CD3<sup>+</sup>CD20<sup>−</sup>CD14<sup>−</sup>CD16<sup>−</sup>CD4<sup>−</sup>CD8<sup>+</sup>) were then sorted using a FACS Aria II (BD Biosciences). Mononuclear cells were separately stained with Live/Dead stain, CD3 AF700 (BD Biosciences, 557917, clone SP34-2), CD4 BV650 (Biolegend, 317436, clone OKT4), CD8 APC-Cy7 (BD Biosciences, 557834, clone SK1), CD14 PB (Biolegend, 301828, clone M5E2), CD20 PB (Biolegend, 302328, clone 2H7) and NKG2A PE (Beckman Coulter, IM3291U, clone Z199) to sort aliquots of 50,000 NK cells (live CD3<sup>+</sup>CD20<sup>−</sup>CD14<sup>−</sup>CD4<sup>−</sup>CD8<sup>−</sup>NKG2A<sup>+</sup>).

### RNA sequencing and data analysis

Bulk CD4<sup>+</sup> T cells were sorted from fresh PBMCs before intervention, and on day 3, week 2 and week 4 post-intervention. In brief, RNA from sorted cells was collected and extracted, and the DNA was digested. Libraries were prepared and normalized, pooled and clustered on a flow cell for sequencing. RNA-sequencing data were aligned to the MacaM v.7.8 assembly of the Indian rhesus macaque genome. To identify pathways that were differentially modulated, gene-set enrichment analysis<sup>38</sup> was performed on the ranked transcript lists using 1,000 phenotype permutations and random seeding. Gene sets used included the MSigDB H (hallmark) gene sets<sup>39</sup>. Normalized enrichment scores for select upregulated gene sets were used, for which the normalization was group-specific. A normalized enrichment score cut-off of greater than 1.35 for upregulated gene sets with a false-discovery rate of less than 0.2 was used, as per the recommendations for gene-set enrichment analysis. For the generation for heat maps, colours represent log<sub>2</sub>-transformed library-size-normalized read counts scaled to unit variance across transcript vectors and normalized to the baseline median sample value of each transcript.

### Single-genome PCR amplification of SIV<sub>mac239</sub> env sequences

cDNA synthesis and 384-well single-genome PCR amplification (SGA) were performed using an approach similar to previously described

methods<sup>40–42</sup>. In brief, RNA was extracted from cryopreserved plasma samples using the QIAmp viral RNA kit (Qiagen, 52906) and reverse transcription was performed using the SuperScript III kit (Invitrogen, 18080-044) with reverse primer SM-ER1 (5'-CTATCACTGTAATAAATCCCTTCCAGTCCC-3'). cDNA was diluted to result in less than 30% positive wells for SGA. First-round PCR was performed in a 15-μl volume using the Phusion Hotstart II High Fidelity DNA Polymerase (Thermo Scientific, F537S) with forward primer H2SM-EF1 (5'-CCCTTGAAGGMGCMRGAGAGCTCATTA-3') and SM-ER1. Cycling conditions were as follows: 98 °C for 2 min; 10 cycles of 95 °C for 15 s, 54 °C for 60 s and 68 °C for 4 min; 25 cycles of 95 °C for 15 s, 54 °C for 60 s and 68 °C for 4 min, adding 5 s to the extension per cycle; 72 °C for 30 min; and 4 °C hold. Second-round PCR was performed with the same enzyme in a 10-μl volume with 1 μl of the first-round PCR reaction as template and primers H2SM-EF2 (5'-CACCTAAAAARTGYTGCTATGCTTCCAG-3') and SM-ER2 (5'-ATAAAATGAGACATGCTATTGCCAATTG-3'). Cycling conditions were as follows: 95 °C for 2 min; 30 cycles of 95 °C for 15 s, 54 °C for 60 s and 72 °C for 2.5 min; 72 °C for 10 min; and 4 °C hold. PCR amplicons were purified using a Qiaquick PCR Purification Kit (Qiagen 28106).

### Sequencing of env amplicons

On average, 26 SGA PCR amplicons per time point (range, 20–30) were sequenced by Eurofins Genomic DNA Sanger sequencing using the following primers: SIVmac251seqF1 5'-GGGATATGTTATGAGCAGT-CACG-3'; SIVmac251seqF2 5'-ATCCAAGAGTCTTGTGACAAGC-3'; SIVmac251seqF3 5'-AAGAGAGGGAGACCTCACG-3'; SIVmac251seqF4 5'-AGGCCAGTGTCTCTTCC-3'; SIVmac251seqR1 5'-CTTGTTCCAA-GCCTGTGC-3'; SIVmac251seqR2 5'-CCTCTGCAATTTGTCCACATG-3'; SIVmac251seqR3 5'-TCCAAGAAGTCAACCTTTTCG-3'; SIVmac251seqR4 5'-AGCTGGGTTTCTCCATGG-3'<sup>18</sup>. Sequencher v.5.1 was used to generate nucleotide sequence contigs and sequences with mixed peaks in the chromatogram were excluded from further analysis.

### Sequence analysis of SIV<sub>mac239</sub> contigs

Geneious v.9.1.7 was used to translate nucleotide sequences into amino acids and generate alignments. Amino acid alignments were exported from Geneious in FASTA format and used to generate highlighter plots to visualize amino acid mismatches ([http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter\\_top.html](http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html)). Phylogenetic neighbour-joining consensus trees (Jukes–Cantor, resampling with 100 bootstrap replicates) were created in Geneious using amino acid alignments, and were exported in NEXUS format into Figtree v.1.4.4 for further modification (A. Rambaut, <http://tree.bio.ed.ac.uk/>). Phylogenetic trees were presented as unrooted or were rooted on the midpoint. Bootstrap values that were higher than 80% are considered significant. Pairwise differences between the infecting SIV<sub>mac239</sub> clone and each SGA-derived Env amino acid sequence were determined using Geneious.

### Experimental design of the BLT humanized mouse model

The BLT humanized model of HIV infection was used to determine the efficacy of CD8 depletion alone, CD8 depletion in combination with N-803 or N-803 alone as a LRA. BLT mice (15–19 weeks after humanization surgery) were exposed to HIV-1<sub>JR-CSF</sub> intravenously. ART was initiated 4–5 weeks later. Viraemia was durably suppressed by ART for 4 weeks. A single dose of N-803, CD8-depleting antibody or the combination of N-803 and CD8-depleting antibody was administered to HIV-infected and suppressed mice as indicated below. HIV RNA induction was measured on days 4 and 7.

### Generation of BLT humanized mice

BLT humanized mice were prepared as previously reported<sup>43–46</sup>. In brief, a 1–2-mm piece of human liver tissue was sandwiched between two pieces of autologous thymus tissue (Advanced Bioscience Resources) under the kidney capsule of sublethally irradiated (200 cGy)

# Article

12–15-week-old NOD.Cg-Prkdc<sup>scid</sup>/Il2rg<sup>tm1Wjl</sup>/SzJ (NSG; The Jackson Laboratory) mice. After implantation, mice were transplanted intravenously with haematopoietic CD34<sup>+</sup> stem cells isolated from autologous human liver tissue. Human immune cell reconstitution was monitored in the peripheral blood of BLT mice by flow cytometry every 3–4 weeks as previously described<sup>43–46</sup>. Mice were maintained under specific-pathogen-free conditions by the Division of Comparative Medicine at the University of North Carolina, Chapel Hill. Mouse experiments were conducted in accordance with NIH guidelines for the housing and care of laboratory animals and in accordance with protocols reviewed and approved by the IACUC at the University of North Carolina, Chapel Hill.

## Production of HIV and infection of BLT mice

Stocks of HIV-1<sub>JR-CSF</sub> were prepared as previously reported<sup>43–45</sup>. The proviral clone was transfected into human embryonic kidney (HEK)293T cells using Lipofectamine 2000 (Invitrogen, 11668030) following the manufacturer's protocol. Viral supernatants were collected 48 h after transfection and titred onto TZM-bl indicator cells in triplicate to determine the tissue-culture infectious units (TCIU) per ml. At least two different titre determinations were performed for each virus stock. BLT mice were exposed to  $3 \times 10^4$  TCIU HIV-1<sub>JR-CSF</sub> by tail-vein injection.

## Analysis of HIV infection in BLT mice

The peripheral-blood plasma viral load was monitored longitudinally by quantitative real-time PCR using a TaqMan RNA-to-C<sub>1</sub>-step kit (Applied Biosystems, 4392656). The sequences of the forward and reverse primers and the TaqMan probe for PCR amplification and detection of HIV *gag* RNA were: 5'-CATGTTTTCAGCATTATCAGAAGGA-3', 5'-TGCTTGATGTCCCCCACT-3' and 5'-FAM-CCACCCACAAGATTAAACAC-CATGCTAAQ-3', respectively. Known quantities of HIV *gag* RNA were run in parallel, creating a standard curve for HIV *gag* and sample RNA levels were quantified by extrapolation from the standard curve. All samples were run and analysed on an ABI 7500 Fast Real-time PCR System (Applied Biosystems).

HIV DNA levels were measured in tissue cells collected at the time of collection and cryopreserved in cryopreservation medium (10% DMSO:90% fetal bovine serum). Cells were thawed slowly, counted by Trypan exclusion, aliquoted and pelleted. DNA was extracted from cell pellets using the QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacturer's instructions. RT-PCR was performed with a TaqMan Fast Universal PCR Master Mix (Applied Biosystems). The sequences of the forward primer, reverse primer and the TaqMan probe for amplification and detection of HIV *gag* DNA were: 5'-CATGTTTTCAGCATTATCAGAAGGA-3', 5'-TGCTTGATGTCCCCCACT-3' and 5'-FAM-CCACCCACAAGATTAAACACCATGCTAAQ-3', respectively. As a control, *Homo sapiens* haemoglobin subunit gamma-2 (*HBG2*) was included to quantify the presence of human DNA in each sample. The sequences of the forward primer, reverse primer and the TaqMan probe for amplification and detection of *HBG2* were 5'-CGCTTCTGGAACGTCTGAGATT-3', 5'-CCTTGTCTCTCTGTGAAATGA-3' and 5'-FAM-TCAATAAGCTCCTAGTCCAGAC-3', respectively. All samples were run and analysed on an ABI 7500 Fast Real-time PCR System (Applied Biosystems).

## ART administration in BLT mice

ART was administered to BLT humanized mice as previously described<sup>47–49</sup>, via 12.7-mm pellets of irradiated Teklad chow containing emtricitabine (1,500 mg kg<sup>-1</sup>), tenofovir disoproxil fumarate (1,560 mg kg<sup>-1</sup>) and raltegravir (600 mg kg<sup>-1</sup>) (Research Diets).

## N-803 and MT807R1 administration in BLT mice

N-803 (0.2 mg kg<sup>-1</sup> in PBS) and the control vehicle (PBS) were administered to mice intravenously in a total volume of 200 µl. MT807R1 (3 mg kg<sup>-1</sup> in PBS) and the control vehicle (PBS) were also administered intravenously in a total volume of 200 µl.

## Immunophenotypic analysis of BLT mice

Immunophenotyping was performed on peripheral-blood samples longitudinally and at the time of collection on blood and mononuclear cells isolated from the tissues of BLT mice. All flow cytometry data were collected on a BD FACSCanto instrument using BD FACSDiva software (v.6.1.3) and data were analysed with FlowJo Software (v.10.5.0). Antibodies for the analysis of human immune cell levels included: CD45 APC (BD 555485, clone HI30), CD3 FITC (BD Biosciences, 555339, clone HIT3a), CD4 APC-Cy7 (BD Biosciences, 560158, clone RPA-T4), CD33 PE (BD Biosciences, 340679, clone P67.6); CD19 PE-Cy7 (BD Biosciences, 557835, clone SJ25C1) and CD8 PerCP (BD Biosciences, 347314, clone SK1). Flow cytometry gating for the expression of lineage-specific antigens on human leukocytes was performed as follows. Step 1, forward and side scatter were used to set a live-cell gate. Step 2, live cells were then analysed for the expression of the human pan-leukocyte marker CD45. Step 3, human leukocytes were then analysed for human CD3<sup>+</sup> T cells and CD19<sup>+</sup> B cells. Step 4, T cells were analysed for human CD4 and CD8 expression. The following flow cytometry antibody panel was also used to analyse HLA-DR, CD38 and CD25 expression: CD3 BV421 (BD Biosciences, 562426, clone UCHL1), CD4 BV605 (BD Biosciences, 562658, clone RPA-T4), CD45 FITC (BD Biosciences, 347463, clone 2D1), HLA-DR PerCP (BD Biosciences, 347364, clone L243), CD69 PE (BD Biosciences, 555531, clone FN50), anti-CD38 PE-Cy7 (BD Biosciences, 335790, clone HB7), CD25 APC (BD Biosciences, 340938, clone 2A3), CD8 APC-Cy7 (BD Biosciences, 557834, clone SK1) and AQUA (ThermoFisher, L35957). Flow cytometry gating was performed as follows. Step 1, forward-scatter height and forward-scatter area were used to eliminate doublets. Step 2, side-scatter area and forward-scatter area were used to distinguish leukocytes based on morphology. Step 3, the viability dye AQUA was used to discriminate live cells from dead cells. Step 4, live cells were analysed for the expression of the human pan-leukocyte marker CD45. Step 5, human leukocytes were then assessed for human CD3 expression to identify T cells. Step 6, T cells were evaluated for the expression of human CD4 and CD8. Step 7, human CD4<sup>+</sup> or CD8<sup>+</sup> T cells were examined for expression of HLA-DR and/or CD38, or CD25. Gates were set with fluorescence-minus-one controls. Non-specific binding was assessed with isotype controls.

## Experimental design CD8 in vitro suppression assay

In vitro latently infected memory CD4<sup>+</sup> T cells were generated using the LARA method as previously described<sup>32</sup> with the following modifications. On day 0, after PBMCs were isolated from buffy coats of HIV-naïve individuals (New York Blood Center) using SepMate density-gradient centrifugation (StemCell Technologies, 85460), a portion of PBMCs from each HIV-naïve donor was cryopreserved in fetal bovine serum (VWR Life Science Seradigm, 97068-085) and 10% DMSO, and stored in liquid nitrogen. On day 8, PBMCs were thawed and rested overnight in RPMI 1640 medium (Fisher Scientific, SH3002701.01) supplemented with 10% fetal bovine serum, 1% penicillin–streptomycin (Corning, 45000-650), and 1% HEPES (Gibco, 15630080); cRPMI before total CD8<sup>+</sup> T-cell-positive enrichment on day 9 (Miltenyi, 130-045-201). CD8<sup>+</sup> T cells were stimulated with anti-CD3/CD28 beads (Dynabeads, 11141D) at a 1:1 ratio plus 30 U ml<sup>-1</sup> IL-2 (R&D Systems, 2021L050CF) for 3 days. On day 12 of LARA, CD4<sup>+</sup> and CD8<sup>+</sup> were prepared for co-culture. HIV latently infected memory CD4<sup>+</sup> T cells were washed, counted and plated for latency reversal in cRPMI in the presence of ART (100 nM efavirenz, 200 nM raltegravir and 5 µM saquinavir). Activated total CD8<sup>+</sup> T cells were removed from the anti-CD3/CD28 beads, washed and resuspended in cRPMI plus ART. CD4<sup>+</sup> and CD8<sup>+</sup> T cells were co-cultured at a 1:1 or 1:5 ratio at a final density of  $1 \times 10^6$  cells per ml. CD4<sup>+</sup> monocultures were also maintained in parallel. Mono- and co-cultures were then left unstimulated, TCR-activated with 1 µg ml<sup>-1</sup> plate-bound OKT3 and 1 µg ml<sup>-1</sup> soluble CD28 (Biolegend, 302933), or treated with 14 nM N-803 (provided by NantKwest) or 500 ng ml<sup>-1</sup> IL-15 (R&D Systems, 247-ILB).

Cells were collected after 72 h and analysed by flow cytometry and qPCR.

Multicolour flow analysis of cell-surface and intracellular marker expression was performed with a BD FACSymphony flow cytometer. Between 200,000 and 600,000 events were acquired for each sample using the live-cell gate. The data were analysed with FlowJo (v.10).

Antibodies used in in vitro suppression assays include: CD3 AF700 (BD Biosciences, 557943, clone UCHT1), CD8 BV737 (BD Biosciences, 564629, clone SK1), HIV-1 core antigen-FITC (Beckman Coulter, 6604665, clone KC57), CD4 BV421 (BD Horizon, 565997, clone SK3), CD45RA APC-eFluor780 (Invitrogen, 47045842, clone HI100), CD27 BV650 (Biolegend, 302828, clone O323), CCR7 Pe-Cy7 (BD Pharmingen, 557648, clone 3D12) and Fixable Viability Dye eFluor 506 (Invitrogen eBioscience, 65-0866-18).

## Statistics and reproducibility

Statistical analyses, including two-way Kruskal–Wallis tests, two-way Friedman tests, one-way ANOVA and Spearman *r* correlations, were performed using GraphPad Prism v.7.0 or v.8.0. Data are mean  $\pm$  s.e.m. as indicated.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Illumina sequencing reads for RNA-sequencing experiments were submitted to the NCBI SRA repository (accession number SRP188630). RNA-sequencing datasets were submitted to the NCBI GEO repository (accession number GSE128415). *env* nucleotide sequences have been deposited in the GenBank (accession numbers MK922999–MK923550).

34. Reed, L. J. & Muench, H. A simple method of estimating fifty per cent endpoints. *Am. J. Epidemiol.* **27**, 493–497 (1938).
35. Chahroudi, A. et al. Target cell availability, rather than breast milk factors, dictates mother-to-infant transmission of SIV in sooty mangabeys and rhesus macaques. *PLoS Pathog.* **10**, e1003958 (2014).
36. Kumar, N. A. et al. Antibody-mediated CD4 depletion induces homeostatic CD4<sup>+</sup> T cell proliferation without detectable virus reactivation in antiretroviral therapy-treated simian immunodeficiency virus-infected macaques. *J. Virol.* **92**, e01235-18 (2018).
37. Deleage, C. et al. Defining HIV and SIV reservoirs in lymphoid tissues. *Pathog. Immun.* **1**, 68–106 (2016).
38. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
39. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
40. Smith, S. A. et al. Diversification in the HIV-1 envelope hyper-variable domains V2, V4, and V5 and higher probability of transmitted/founder envelope glycosylation favor the development of heterologous neutralization breadth. *PLoS Pathog.* **12**, e1005989 (2016).
41. Smith, S. A. et al. Signatures in simian immunodeficiency virus SIVsmE660 envelope gp120 are associated with mucosal transmission but not vaccination breakthrough in rhesus macaques. *J. Virol.* **90**, 1880–1887 (2016).
42. Burton, S. L. et al. Breakthrough of SIV strain smE660 challenge in SIV strain mac239-vaccinated rhesus macaques despite potent autologous neutralizing antibody responses. *Proc. Natl Acad. Sci. USA* **112**, 10780–10785 (2015).
43. Denton, P. W. et al. Antiretroviral pre-exposure prophylaxis prevents vaginal transmission of HIV-1 in humanized BLT mice. *PLoS Medicine* **5**, e16 (2008).
44. Denton, P. W. et al. Generation of HIV latency in humanized BLT mice. *J. Virol.* **86**, 630–634 (2011).
45. Denton, P. W. et al. One percent tenofovir applied topically to humanized BLT mice and used according to the CAPRISA 004 experimental design demonstrates partial protection from vaginal HIV infection, validating the BLT model for evaluation of new microbicide candidates. *J. Virol.* **85**, 7582–7593 (2011).
46. Melkus, M. W. et al. Humanized mice mount specific adaptive and innate immune responses to EBV and TSST-1. *Nat. Med.* **12**, 1316–1322 (2006).
47. Honeycutt, J. B. et al. T cells establish and maintain CNS viral infection in HIV-infected humanized mice. *J. Clin. Invest.* **128**, 2862–2876 (2018).
48. Kessing, C. F. et al. In vivo suppression of HIV rebound by didehydro-cortistatin A, a “block-and-lock” strategy for HIV-1 treatment. *Cell Rep.* **21**, 600–611 (2017).
49. Tsai, P. et al. In vivo analysis of the effect of panobinostat on cell-associated HIV RNA and DNA levels and latent HIV infection. *Retrovirology* **13**, 36 (2016).

**Acknowledgements** This work was supported by NIH grants R01-AI125064 and UM1-AI124436 (to G.S. and A.C.); R01-AI143414 (to G.S. and D.A.K.); R01-MH108179 and R01-AI111899 (to J.V.G.); UM1-AI126619 (to D.M.M.); R01-AI123010 (to A.W.); P30 AI050409 (Emory Center for AIDS Research); P51 OD011092 (Oregon National Primate Research Center base grant); the National Institutes of Health's Office of the Director, Office of Research Infrastructure Programs P51OD011132 (Yerkes National Primate Research Center base grant); and supported in part by federal funds from the National Cancer Institute, National Institutes of Health, under contracts HHSN261200800001E and 75N91019D00024 (J.D.L.). We thank B. Jones, S. O'Connor and J. Sacha for discussions; S. Ehnert, S. Jean and all of the animal care and veterinary staff at the Yerkes National Primate Research Center; B. Cervasi and K. Gill at the Emory University Flow Cytometry Core; Emory and Pediatric's/Winship Flow Cytometry Core; the Translational Virology and Reservoir Cores of the Emory CFAR, the Emory Nonhuman Primate Genomics Core for RNA sequencing and analysis and the Quantitative Molecular Diagnostics Core of the AIDS and Cancer Virus Program, Frederick National Laboratory, for high-sensitivity plasma viral-load testing; NantKwest for providing N-803, K. Reimann and the NHP Reagent Resources for the MT807R1 antibody, R. Geleziunas and Gilead Pharmaceuticals for providing tenofovir and emtricitabine, D. Hazuda and B. Howell from Merck for providing raltegravir and J. Demarest and Viiv Healthcare for providing dolutegravir for this study.

**Author contributions** J.B.M., A.C., M.P. and G.S. designed the experiments. J.B.M., M.M., E.W. and D.G.C. performed the experiments. S.A.S. performed single-genome PCR and sequencing of SIV RNA. S.A.S. and C.A.D. performed the sequence-based analyses and wrote the relevant portions of the manuscript. J.D.L. performed ultrasensitive viral-load analyses. B.C. performed FACS of live cells. T.H.V. measured viral load, and cell-associated DNA and RNA. J.D.E. and K.B.-S. performed RNAscope analysis. S.E.B., G.K.T. and H.W. performed RNA-sequencing analysis. M.K., C.R.A.-S. and W.O.T. constructed BLT humanized mice. M.K. performed the HIV infection and ART suppression of BLT humanized mice. R.A.S. performed the viral load measurements, the isolation of nucleic acids and the analysis of tissue RNA levels for BLT humanized mice. C.R.A.-S. and W.O.T. designed and performed the N-803 and CD8 T cell depletion experiments in BLT humanized mice and analysed the data. A.W. supervised the data collection, analysis, figure preparation and reporting of all samples from the BLT humanized mice. J.V.G. designed, coordinated and supervised all of the BLT experimental work. C.R.A.-S., J.V.G., A.W. and W.O.T. wrote and revised the BLT humanized mice portions of the manuscript. L.F. and D.A.K. performed the in vitro studies and wrote the relevant portions of the manuscript. D.M.M., J.T.S. and J.H.L. provided technical support. J.B.M., A.C. and G.S. wrote the manuscript.

**Competing interests** The authors declare no competing interests.

## Additional information

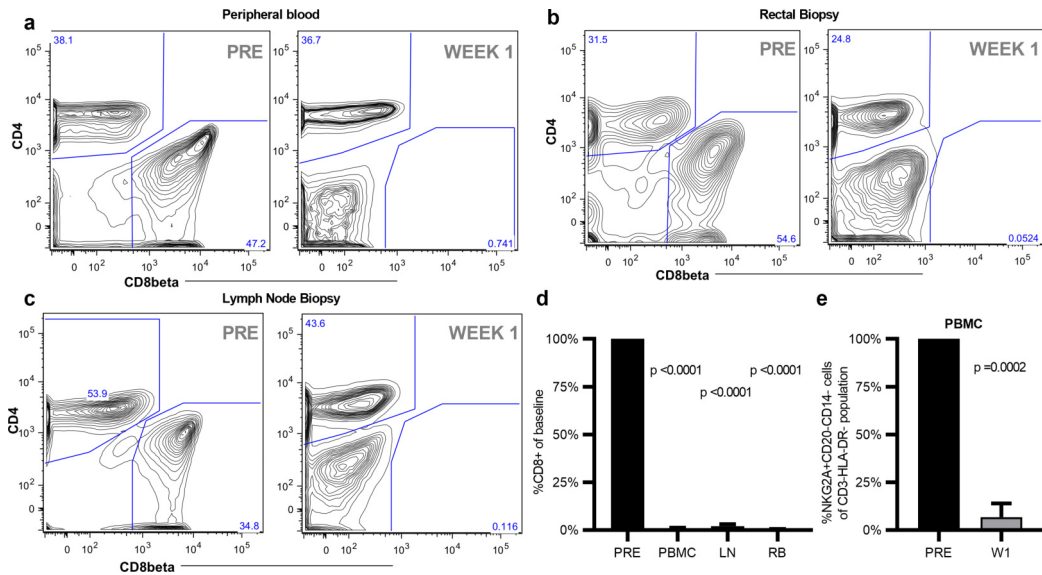
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1946-0>.

**Correspondence and requests for materials** should be addressed to G.S.

**Peer review information** Nature thanks Mathias Lichterfeld and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

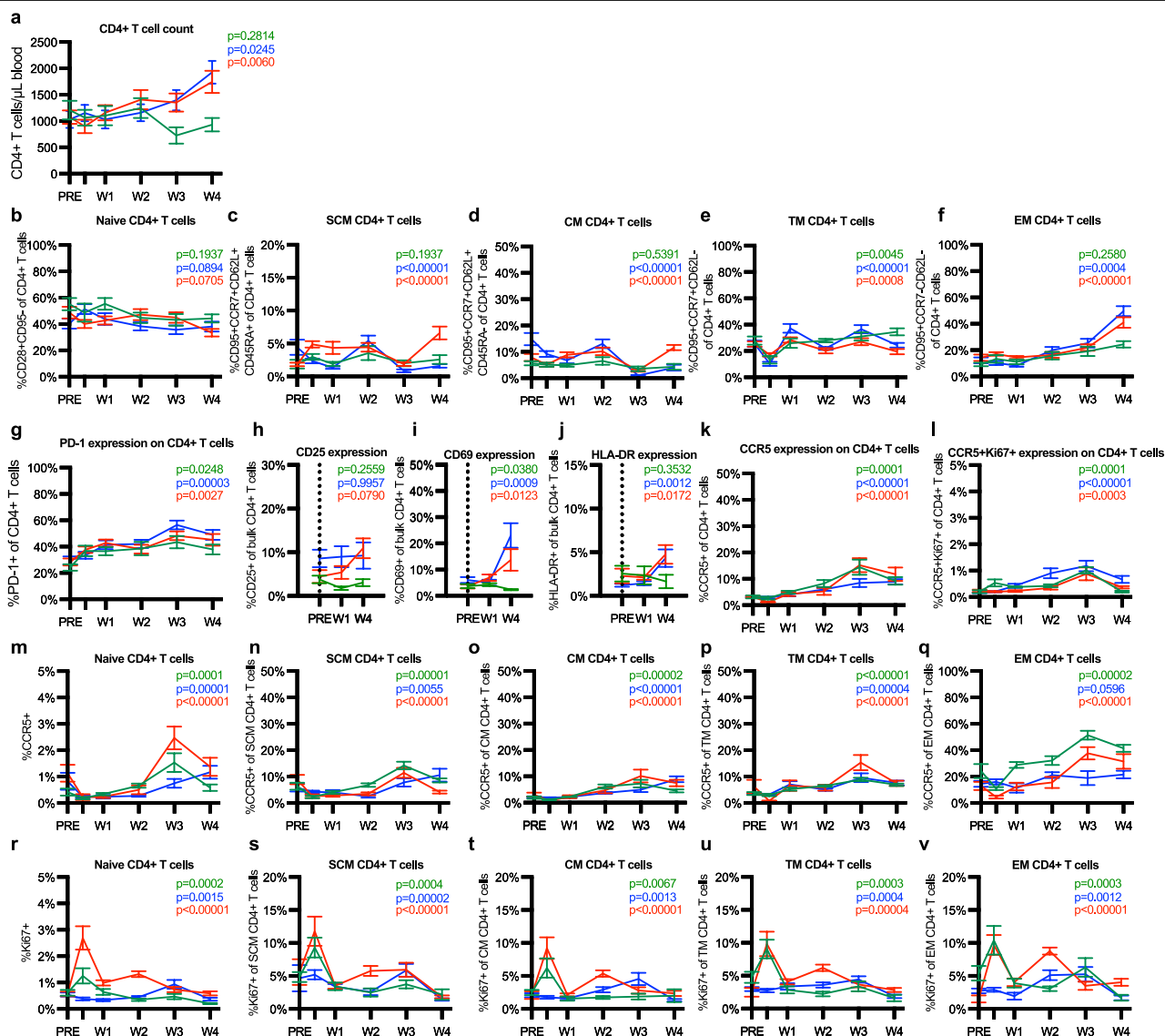
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





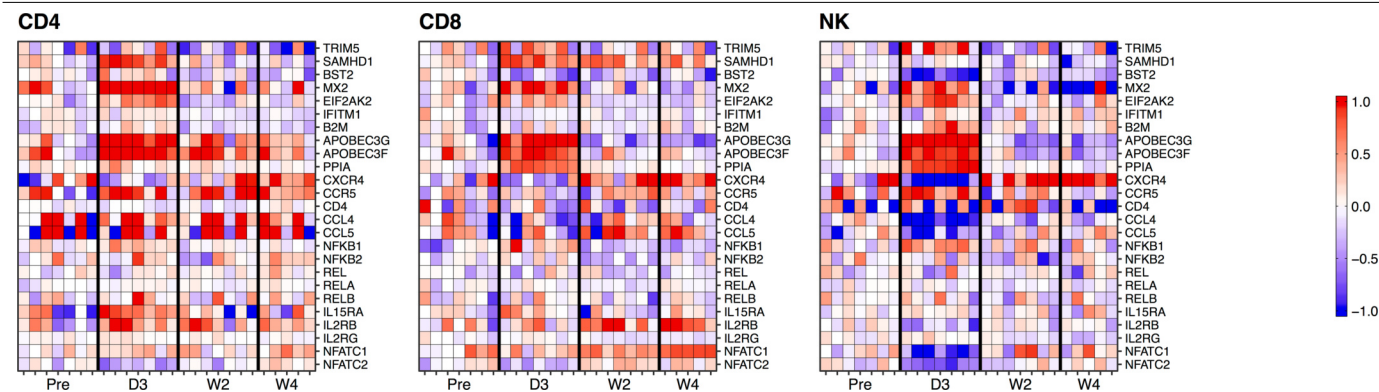
**Extended Data Fig. 1 | MT807R1 effectively depletes CD8<sup>+</sup> T cells in peripheral blood, lymph node and rectum of macaques in addition to depleting NK cells from the blood at day 7.** The percentage of CD8<sup>+</sup> cells in the CD3<sup>+</sup> population 7 days after depletion was compared to the levels before depletion. **a–c**, Sample flow cytometry shows the absence of CD8β<sup>+</sup> cells as part of the CD3<sup>+</sup> T cell population after depletion in the peripheral blood (**a**), rectum (**b**) and lymph nodes (**c**) of macaques. Similar results were found across all CD8-depleted macaques ( $n = 28$  biologically independent samples). **d**, The percentage of CD8β<sup>+</sup> cells compared to pre-depletion baseline was calculated

in all CD8-depleted macaques (treated with or without N-803,  $n = 28$  macaques) across blood and tissue samples (no differences in CD8<sup>+</sup> T cell depletion were observed between groups 2 and 3 on day 7). A two-sided Friedman test was used to calculate statistically significant changes compared to baseline across tissues. **e**, Depletion of NK cells in the peripheral blood was assessed 1 week after CD8 depletion alone ( $n = 14$  macaques) compared to baseline. Statistical significance was calculated using Wilcoxon signed-rank test. Data are mean  $\pm$  s.e.m.



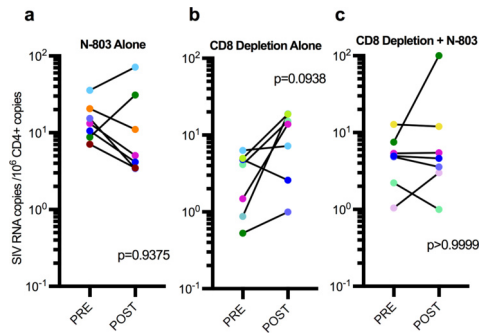
**Extended Data Fig. 2 | Phenotypic changes to CD4<sup>+</sup> T cells after intervention.** Longitudinal flow cytometry analysis after treatment with only N-803 (green,  $n = 7$  macaques), only CD8 depletion (blue,  $n = 14$  macaques) and after CD8 depletion and treatment with N-803 (red,  $n = 14$  macaques). **a**, CD4<sup>+</sup> T cell frequency. **b–f**, Percentage of naive (b), stem cell memory (SCM) (c), central memory (CM) (d), transitional memory (TM) (e) and effector memory (EM) (f)

CD4<sup>+</sup> T cells. **g–l**, Percentage of bulk CD4<sup>+</sup> T cells that express PD-1 (g), CD25 (h), CD69 (i), HLA-DR (j), CCR5 (k) and both CCR5 and Ki-67 (l). **m–v**, CCR5 (m–q) and Ki-67 (r–v) expression in different subsets of CD4<sup>+</sup> T cells. Data are mean  $\pm$  s.e.m. Two-sided Kruskal–Wallis tests were used to compare values after the intervention to the pre-intervention baseline and approximate  $P$ -value summaries are provided.

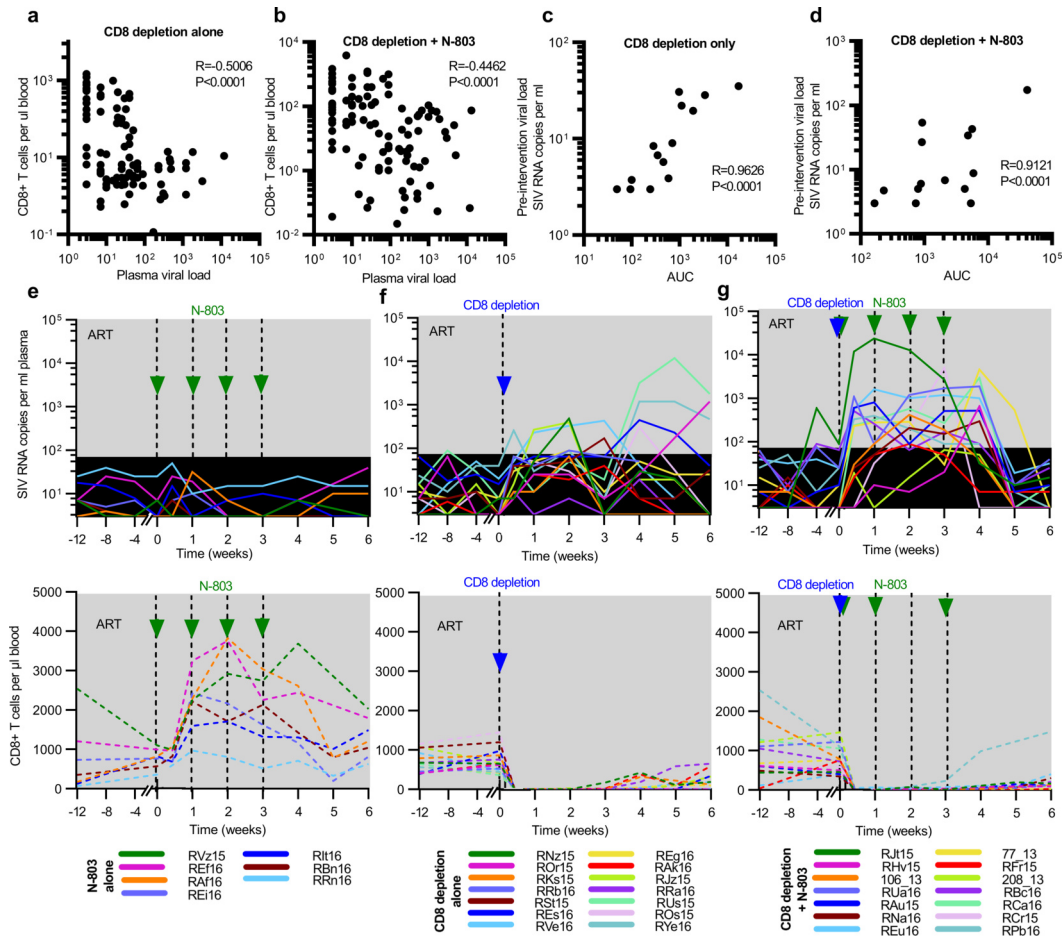


**Extended Data Fig. 3 | SIV-associated genes and IL-15 subunit genes show a transient change in expression after treatment with N-803 alone.** RNA was extracted from sorted peripheral bulk CD4<sup>+</sup> T cells (CD3<sup>+</sup>CD4<sup>+</sup>CD8<sup>-</sup>CD20<sup>-</sup>CD14<sup>-</sup>) (left), bulk CD8<sup>+</sup> T cells (CD3<sup>+</sup>CD4<sup>-</sup>CD8<sup>+</sup>CD20<sup>-</sup>CD14<sup>-</sup>) (middle) and NK cells (CD3<sup>+</sup>CD20<sup>-</sup>CD14<sup>-</sup>CD16<sup>+</sup>) (right) and libraries were prepared, normalized, pooled and clustered on flow cells for sequencing. RNA-

sequencing data were aligned to the MacaM v.7.8 assembly of the Indian rhesus macaque genome. Transcripts were analysed for alignment against a custom gene set with SIV host restriction factors, PPIA (capsid folding protein), SIV receptors, SIV receptor agonists, NF- $\kappa$ B subunits (involved in mediating the transcription of long-terminal repeat regions), IL-15 receptor subunits and NFAT subunits. D, day; W, week.



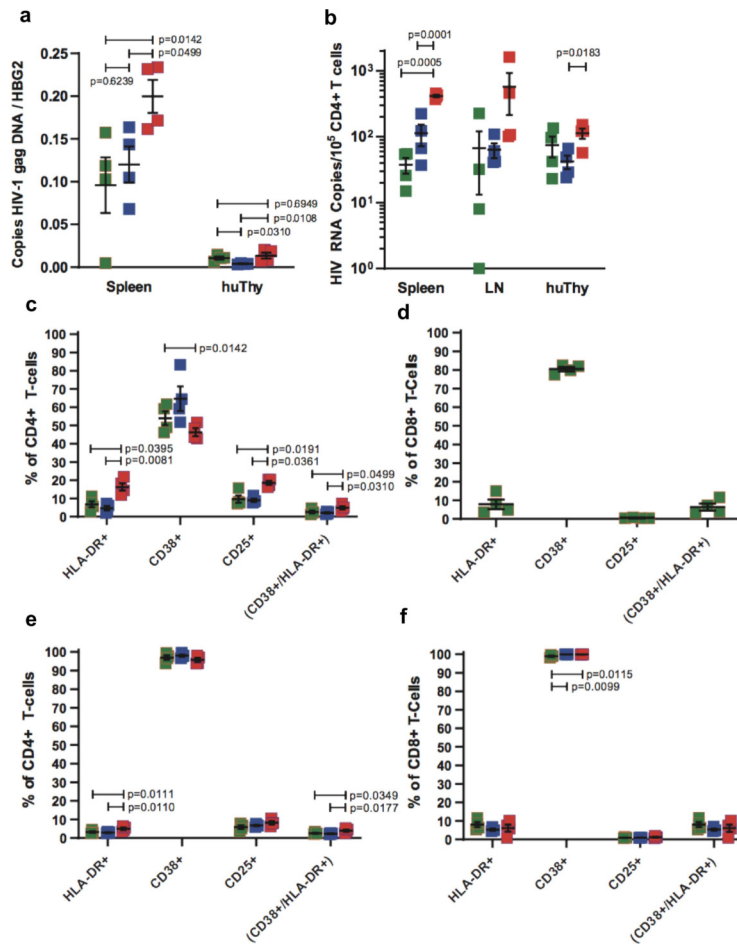
**Extended Data Fig. 4 | Quantification of levels of cell-associated SIV RNA in peripheral CD4<sup>+</sup> T cells before and after interventions.** a–c, Changes in expression of SIV RNA in relation to the number of copies of CD4 after intervention with N-803 alone (*n* = 7 macaques) (a), CD8 depletion alone (*n* = 7 macaques) (b) or CD8 depletion combined with N-803 (*n* = 7 macaques) (c). Data are mean ± s.e.m. Two-sided Wilcoxon signed-rank tests were used to compare values after the intervention to the pre-intervention baseline.



**Extended Data Fig. 5 | Level of virus reactivation correlated with the absence of CD8<sup>+</sup> T cells.** **a, b**, Correlation between CD8<sup>+</sup> T cell counts and viral load (SIV RNA copies per ml of plasma) on day 0, day 3, and weekly up to week 6 after interventions. **a**, CD8 depletion alone ( $n = 103$  samples from 14 macaques). **b**, CD8 depletion combined with N-803 treatment ( $n = 112$  samples from 14 macaques). **c, d**, The area under the curve (AUC) and the average pre-intervention viral load after CD8 depletion alone (**c**;  $n = 14$  macaques) or CD8

depletion with N-803 treatment (**d**;  $n = 14$  macaques). Correlation coefficients are calculated using the Spearman's rank-order correlation (two-tailed, no adjustments). **e–g**, Longitudinal viral loads (top) and CD8<sup>+</sup> T cell counts (bottom) after N-803 treatment alone (**e**;  $n = 7$  macaques), CD8 depletion alone (**f**;  $n = 14$  macaques) or CD8 depletion combined with N-803 treatment (**g**;  $n = 14$  macaques). Colour keys along the bottom indicate animal IDs.





**Extended Data Fig. 6 | HIV DNA, HIV RNA and human T cell activation levels in HIV-infected, ART-suppressed BLT humanized mice treated with N-803, CD8 depletion alone or combined CD8 depletion with N-803.** **a, b,** HIV-infected, ART-suppressed BLT humanized mice were treated with N-803 (green,  $n = 4$  BLT humanized mice), CD8 depletion alone (blue,  $n = 4$ ), or treated with CD8 depletion with N-803 (red,  $n = 4$ ). After 7 days, total HIV DNA (**a**) and cell-associated HIV RNA (**b**) were extracted from mononuclear cells isolated from the spleen, human-derived thymus (huThy) and lymph node (LN; HIV RNA

only). **c-f,** Percentages of HLA-DR<sup>+</sup>, CD38<sup>+</sup>, CD25<sup>+</sup> or HLA-DR<sup>+</sup>CD38<sup>+</sup> cells were measured in human CD4<sup>+</sup> (**c, e**) or CD8<sup>+</sup> (**d, f**) T cells isolated from the spleen (**c, d**) or human-derived thymus (**e, f**) of HIV-infected, ART-suppressed BLT humanized mice 7 days after treatment with N-803 (green,  $n = 4$ ), CD8 depletion alone (blue,  $n = 4$ ) or CD8 depletion combined with N-803 treatment (red,  $n = 4$ ). Treatment groups were compared using a two-tailed Student's *t*-test (**a**) or a Kruskal–Wallis test with a false-discovery rate correction (**b–f**). Data are mean  $\pm$  s.e.m.

77\_13

RAu15

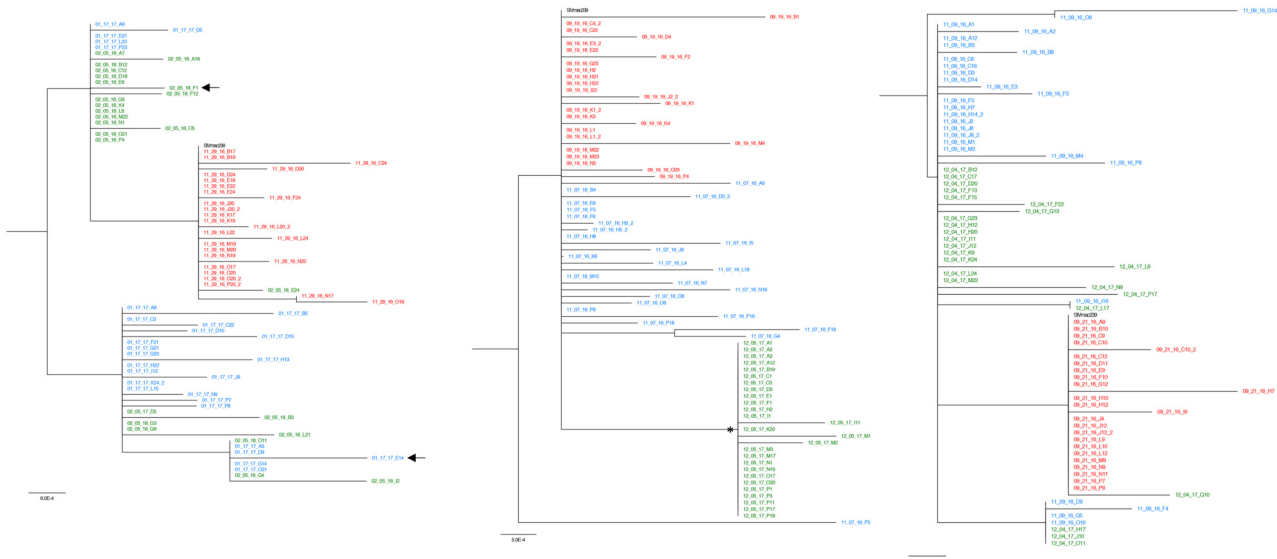
RCr15



REu16

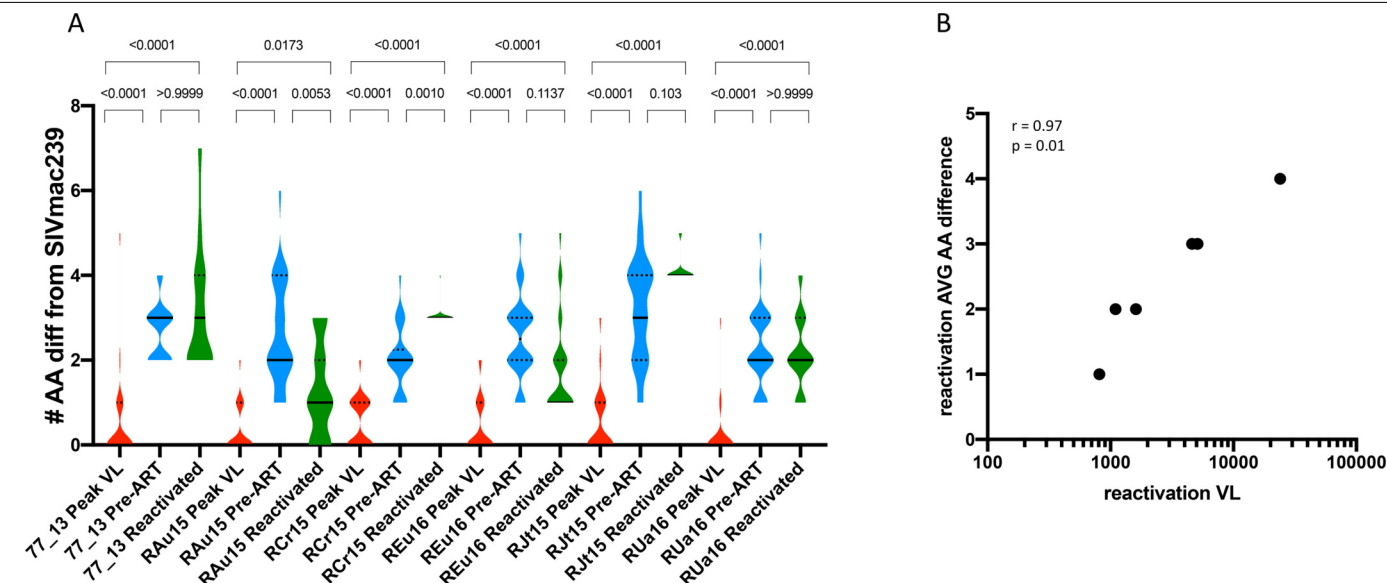
RJt15

RUa16



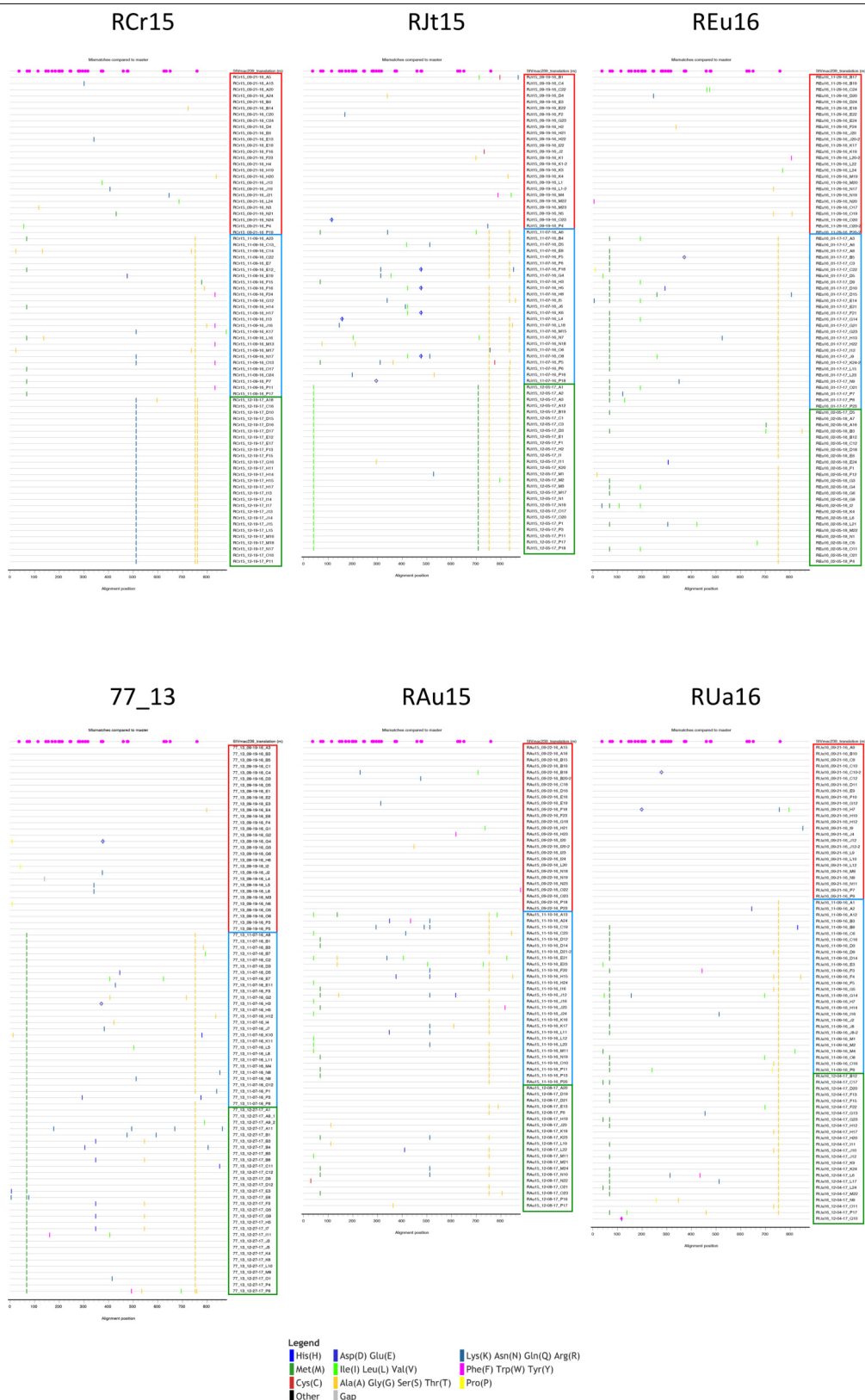
**Extended Data Fig. 7 | Phylogenetic trees of longitudinal SGA-derived Env amino acid sequences.** Phylogenetic trees were generated for six macaques that received CD8 depletion with N-803 using Env sequences from the peak viral load (red), pre-ART (blue) and reactivation (green) time points. The Env sequence of the SIV<sub>mac239</sub> clone used for infection is included in each tree

(black). The horizontal bar below each tree indicates the genetic distance. Sequence clusters that are supported by bootstraps greater than 80% are indicated by an asterisk. Env sequences that contain a stop codon are indicated by an arrow.



**Extended Data Fig. 8 | Longitudinal Env amino acid divergence from the input virus and relationship with viral load in the plasma.** The number of amino acid differences between the infecting viral clone SIV<sub>mac239</sub> and each SGA amplicon was determined using Geneious. **a**, Violin plots show the frequency distribution of the number of amino acid differences between sequences at each time point in each macaque. The solid line indicates the median number of amino acid differences for each individual Env sequence; the dotted lines indicate the quartiles. Peak viral load (VL) (red), pre-ART (blue) and

reactivation (green) time points are shown. The animal ID and the three time points are indicated on the x axis. Statistical differences between time points for each macaque were determined by performing multiple comparisons using a Kruskal–Wallis test with Dunn's correction. **b**, The average number of sequence differences for each macaque at the reactivation time point is plotted on the y axis, and the corresponding plasma viral loads are plotted on the x axis on a log<sub>10</sub> scale. Correlation coefficients are calculated using the Spearman's rank-order correlation (two-tailed, no adjustments).



**Extended Data Fig. 9 | Highlighter plots of longitudinal SGA-derived Env amino acid sequences.** Highlighter plots were generated for six representative macaques that were depleted of CD8 and treated with N-803 using Env sequences from peak VL (red box), pre-ART (blue box) and reactivation (green box) time points. The Env sequence of the SIV<sub>mac239</sub> clone used for infection is

included as the master (reference) sequence in each plot. The position of *N*-linked glycosylation sites on the master sequence are indicated by pink circles. Each tick represents an amino acid difference from the master sequence, as indicated in the key. Blue diamonds indicate the loss of an *N*-linked glycosylation site.

**Extended Data Table 1 | Viral loads from macaque and humanized mouse studies**

Model: ART-treated, SIV-infected rhesus macaques (limit of detection = 3 copies of SIV RNA/mL of plasma)													
		Pre-intervention				Post-intervention							
Intervention	Macaque	Month -3	Month -2	Month -1	Day 0	Day 3	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	
N-803 alone	REf16 R	7	< 3	< 3	7	< 3	< 3	< 3	< 3	7		< 3	
	RVz15 R	7	25	19	7	25	19	< 3	< 3	1600		40	
	RAf16 R	< 3	4	< 3	< 3	< 3	32	7	< 3	< 3	10	10	
	REi16 R	7	5	7	< 3	7	10	< 3	< 3	< 3	< 3	< 3	
	RIi16 R	18	15	8	< 3	15	< 3	7	10	7	< 3	< 3	
	RBn16 R	7	4	< 3	< 3	7	< 3	< 3	< 3	< 3	< 3	< 3	
	RRn16 R	25	40	25	25	50	10	15	15	25	15	15	
CD8 depletion alone	RNz15 R	< 3	< 3	< 3	7	7	50	490	< 3	30	20	< 3	
	ROr15 R	7	65	< 3	< 3	7	25	25	7	25		1200	
	RKs15 R	< 3	< 3	< 3	< 3	10	10	65	< 3	< 3		< 3	
	RRb16 R				< 3	65	40	90	65	50	Nx		
	REs16 R	65	18	25	15	40	65	65	65	450	230	40	
	RS15 R	7	7	10	< 3	40	40	50	170	7	7	32	
	RVe16 R	3	15	50	20	50	230	330	430	40	7	32	
	REg16 R	10	7	10	7	30	50	90	10	50	25	25	
	RAk16R	< 3	6	< 3	< 3	7	25	19	40	7	< 3	< 3	
	RJz15 R	15	< 3	15	< 3	15	270	400	< 3	19	19	< 3	
	RRa16 R	< 3	< 3	< 3	< 3	10	< 3	7	< 3	19	< 3	< 3	
	ROs15 R	7	< 3	10	< 3	65	50	30	7	310	32	< 3	
	RUs15 R	15	90	29	7	40	10	50	19	3200	12000	1800	
	RYe16 R	24	10	40	40	260	30	90	7	1200	1200	470	
	RJ15 R	< 3	< 3	600	90	12000	24000	13000	2700	32	10	15	
	RHv15 R	< 3	< 3	< 3	< 3	< 3	10	7	20	660	< 3	< 3	
	106_13 R	7	7	< 3	< 3	30	90	420	190	40	< 3	< 3	
	RUa16 R	65	7	220	25	1100	90	1200	1700	1900	10	40	
	RAu15 R	7	< 3	7	10	610	810	90	510	520	7	10	
	RNa16 R	< 3	15	< 3	< 3	25	50	210	150	300	< 3	< 3	
	REu16 R	40	130	40	25	557	1610	1000	1200	1000	19	32	
	77_13 R	15	7	< 3	10	230	280	290	30	4600	540	< 3	
CD8 depletion with N-803	RFr15 R	< 3	10	< 3	< 3	15	50	90	50	7	7	7	
	208_13 R	< 3	< 3	< 3	< 3	19	< 3	15	65	50	< 3	< 3	
	RBc16 R	50	10	90	65	510	260	65	170	90	10	25	
	RCa16 R	7	7	< 3	< 3	240	320	580	270	3000	< 3	10	
	RCr15 R	< 3	< 3	< 3	< 3	32		200	5100	< 3	< 3	10	
	RPb16 R	25	50	7	25	330	400	200	90	90	7	< 3	
Model: ART-treated, SHIV-infected rhesus macaques (limit of detection = 3 copies of SIV RNA/mL of plasma)													
		Pre-intervention				Post-intervention							
Intervention	Macaque	Month -3	Month -2	Month -1	Day 0	Day 3	Week 1	Day 10	Week 2	Week 3	Week 4	Week 5	Week 6
CD8 depletion with N-803	RKm16	10	9	7	3	430	30	470	480	3600	90	3	90
	CB91	65	7	9	3	460	60	660	90	810	50	30	3
	RPp16	10	7	7	3	20	7	10	7	7	25	15	15
	RRI16	3	7	7	3	3	3	19	3	7	65	15	3
	RYr16	15	7	7	3	20	50		50	25	40	3	7
Model: ART-treated, HIV-infected bone marrow-liver-thymus humanized mice (limit of detection = 346 copies of SIV RNA/mL of plasma)													
Intervention	Mouse	Pre-intervention	Day 4	Day 7									
N-803 alone	1	<346	<346	<346									
	2	<346	<346	<346									
	3	<346	<346	<346									
	4	<346	<346	<346									
	5	<346	<346	<346									
	6	<346	<346	<346									
	7	<346	<346	<346									
CD8 depletion alone	1	<346	560	<346									
	2	<346	<346	378									
	3	<346	<346	378									
	4	<346	<346	<346									
	5	<346	<346	<346									
	6	<346	<346	<346									
	7	<346	<346	<346									
	8	<346	<346	<346									
CD8 depletion with N-803	1	<346	1300	1488									
	2	<346	1079	<346									
	3	<346	779	<346									
	4	<346	546	574									
	5	<346	354	<346									
	6	<346	<346	1504									
	7	<346	<346	1981									
	8	<346	<346	<346									

Longitudinal viral loads before and after interventions using the ART-treated, SIV<sub>mac239</sub>-infected rhesus macaque model (limit of detection of three copies of SIV RNA per ml of plasma; corresponding to Fig. 2a–d) (top), the ART-treated, SHIV<sub>SF162P3</sub>-infected rhesus macaque model (limit of detection of 3 copies of SHIV RNA per ml of plasma; corresponding to Fig. 2g) (middle) and the ART-treated, HIV<sub>JR-CSF</sub>-infected BLT humanized mouse model (limit of detection of 346 copies of HIV RNA per ml of plasma; corresponding to Fig. 2i–k) (bottom).



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection FACS Diva V8.01, Sequencher V5.1

Data analysis FlowJo V9.9.6./V10.5.0 (Tree Star); Graph Pad Prism V7.0a (Graphpad software), Gene Set Enrichment Analysis V3.0 (GSEA; Broad Institute), Geneious V9.1.7., Figtree V1.4.4.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Illumina sequencing reads for RNA-Seq experiments were submitted to the NCBI SRA repository and are available at Accession #SRP188630. RNA-Seq datasets were submitted to the NCBI GEO repository and are available at accession number GSE128415. Env nucleotide sequences have been deposited into Genbank under the accession number MK922999-MK923550.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Based on our previous data on SIV-infected ART -treated rhesus macaques, with a sample size of at least 7, we would be able to detect a significant difference between pre- and post-CD8 depletion samples in the level of plasma RNA at the 0.05 significance level with a power of 0.90.
Data exclusions	No data exclusion was applied to this study.
Replication	The use of non-human primates precludes our ability to replicate experiments. Sample sizes were chosen to maximize the likelihood of detecting statistical differences.
Randomization	Age, weight, sex, A01 status, peak post-infection viral load, and time to suppression after ART were all controlled for when allocated animals into experimental groups.
Blinding	No blinding was used in this study.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	Macaque studies: CCR5 APC (3A9), CCR7 FITC (150503), CD3 APC-Cy7 (SP34-2), CD4 BV650 (OKT4), CD8α BV711 (RPA-T8), CD8β PE-Cy5 (SID8BEE), CCR5 APC (3A9), CCR7 FITC (150503), CD45RA Pe-Cy7 (5H9), CD62L PE (SK11), CD95 BV605 (DX2), PD-1 BV421 (EH12.2H7), CD16 BV421 (3G8), CD20 PE-Cy5 (2H7), CD14 PE-Cy7 (M5E2), NKG2A (CD159) PE (Z199), CXCR5 PE-eFluor610 (MU5UBEE), CD28 PE-Cy5.5 (CD28.2), and CD56 FITC (NCAM16.2). Human primary cell model: CD3 Alexa Fluor® 700 (UCHT1, BD Biosciences, #557943), CD8 BUV737 (SK1, BD Horizon™, #564629), HIV-1 core antigen-FITC (KC57, Coulter Clone, #6604665), CD4 BV421 (SK3, BD Horizon™, #565997), CD45RA APC- eFluor™780 (HI100, Invitrogen, #47045842), CD27 BV650 (O323, Biolegend, #302828), CCR7 Pe-Cy7 (3D12, BD Pharmingen™, #557648), Fixable Viability Dye eFluor™ 506 (Invitrogen eBioscience #65-0866-18). Humanized mice model: CD45 APC (clone HIT3a; BD Biosciences #555485), CD3 FITC (clone HIT3a; BD Biosciences #555339), CD4 APC-Cy7 (clone RPA-T4; BD Biosciences #560158), CD33 PE (clone P67.6; BD Biosciences #340679); CD19 PE-Cy7 (clone SJ25C1; BD Biosciences #557835) and CD8 PerCP (clone SK1; BD Biosciences #347314), CD3 BV421 (clone UCHT1; BD Biosciences #562426), CD4 BV605 (clone RPA-T4; BD Biosciences #562658), CD45 FITC (clone 2D1; BD Biosciences #347463), HLA-DR PerCP (clone L243; BD Biosciences #347364), CD69 PE (clone FN50; BD Biosciences #555531), anti-CD38 PE-Cy7 (clone HB7; BD Biosciences #335790), CD25 APC (clone 2A3; BD Biosciences #340938), CD8 APC-Cy7 (clone SK1; BD Biosciences #557834), and AQUA (ThermoFisher #L35957).
Validation	Antibodies used in macaque studies were validated in previous studies and the NIH Nonhuman Primate Reagent Resource. Independently validated in monochromatic titration on primary macaque cells. Antibodies used the humanized mice and human primary cell models were also independently validated in monochromatic titration experiments.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Indian-origin rhesus macaques, male and female bred at the Yerkes National Primate Research center. Animals were roughly four years of age at the start of the study.
Wild animals	NA
Field-collected samples	NA
Ethics oversight	All procedures are approved by the Emory University Institutional Animal Care and Use Committee (IACUC) and animal care facilities are accredited by the U.S. Department of Agriculture (USDA) and the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation	Described in methods pages 30-31.
Instrument	BD LSR II
Software	FACS Diva V8.0.1., FlowJo V9.9.6.
Cell population abundance	Purity was high, as determined by flow cytometry of purified sample post-sort.
Gating strategy	As described in methods pages 30-31.
<input type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

# Systemic HIV and SIV latency reversal via non-canonical NF- $\kappa$ B signalling in vivo

<https://doi.org/10.1038/s41586-020-1951-3>

Received: 12 April 2019

Accepted: 16 December 2019

Published online: 22 January 2020

Christopher C. Nixon<sup>1,2,3,20</sup>, Maud Mavigner<sup>4,20</sup>, Gavin C. Sampey<sup>2,3,5,6</sup>, Alyssa D. Brooks<sup>4</sup>, Rae Ann Spagnuolo<sup>1,2,3</sup>, David M. Irlbeck<sup>6,7</sup>, Cameron Mattingly<sup>4</sup>, Phong T. Ho<sup>1,2,3</sup>, Nils Schoof<sup>4</sup>, Corinne G. Cammon<sup>1,2,3</sup>, Greg K. Tharp<sup>8</sup>, Matthew Kanke<sup>9,10</sup>, Zhang Wang<sup>11</sup>, Rachel A. Cleary<sup>1,2,3</sup>, Amit A. Upadhyay<sup>8</sup>, Chandrav De<sup>1,2,3</sup>, Saintedym R. Wills<sup>2,3,5,6</sup>, Shane D. Falcinelli<sup>2,3,5,12</sup>, Cristin Galarini<sup>6,7</sup>, Hasse Walum<sup>8</sup>, Nathaniel J. Schramm<sup>1,2,3</sup>, Jennifer Deutsch<sup>11</sup>, Jeffrey D. Lifson<sup>13</sup>, Christine M. Fennessey<sup>13</sup>, Brandon F. Keele<sup>13</sup>, Sherrie Jean<sup>8</sup>, Sean Maguire<sup>11</sup>, Baolin Liao<sup>1,2,3,14</sup>, Edward P. Browne<sup>2,3,5</sup>, Robert G. Ferris<sup>6,7</sup>, Jessica H. Brehm<sup>6,7</sup>, David Favre<sup>6,11</sup>, Thomas H. Vanderford<sup>8</sup>, Steven E. Bosinger<sup>8,15</sup>, Corbin D. Jones<sup>9,10</sup>, Jean-Pierre Routy<sup>16,17</sup>, Nancie M. Archin<sup>2,3,5</sup>, David M. Margolis<sup>2,3,5,6,12,18</sup>, Angela Wahl<sup>1,2,3</sup>, Richard M. Dunham<sup>2,3,5,6,7,21\*</sup>, Guido Silvestri<sup>8,15</sup>, Ann Chahroudi<sup>4,8,19,21\*</sup> & J. Victor Garcia<sup>1,2,3,21\*</sup>

Long-lasting, latently infected resting CD4<sup>+</sup> T cells are the greatest obstacle to obtaining a cure for HIV infection, as these cells can persist despite decades of treatment with antiretroviral therapy (ART). Estimates indicate that more than 70 years of continuous, fully suppressive ART are needed to eliminate the HIV reservoir<sup>1</sup>. Alternatively, induction of HIV from its latent state could accelerate the decrease in the reservoir, thus reducing the time to eradication. Previous attempts to reactivate latent HIV in preclinical animal models and in clinical trials have measured HIV induction in the peripheral blood with minimal focus on tissue reservoirs and have had limited effect<sup>2–9</sup>. Here we show that activation of the non-canonical NF- $\kappa$ B signalling pathway by AZD5582 results in the induction of HIV and SIV RNA expression in the blood and tissues of ART-suppressed bone-marrow–liver–thymus (BLT) humanized mice and rhesus macaques infected with HIV and SIV, respectively. Analysis of resting CD4<sup>+</sup> T cells from tissues after AZD5582 treatment revealed increased SIV RNA expression in the lymph nodes of macaques and robust induction of HIV in almost all tissues analysed in humanized mice, including the lymph nodes, thymus, bone marrow, liver and lung. This promising approach to latency reversal—in combination with appropriate tools for systemic clearance of persistent HIV infection—greatly increases opportunities for HIV eradication.

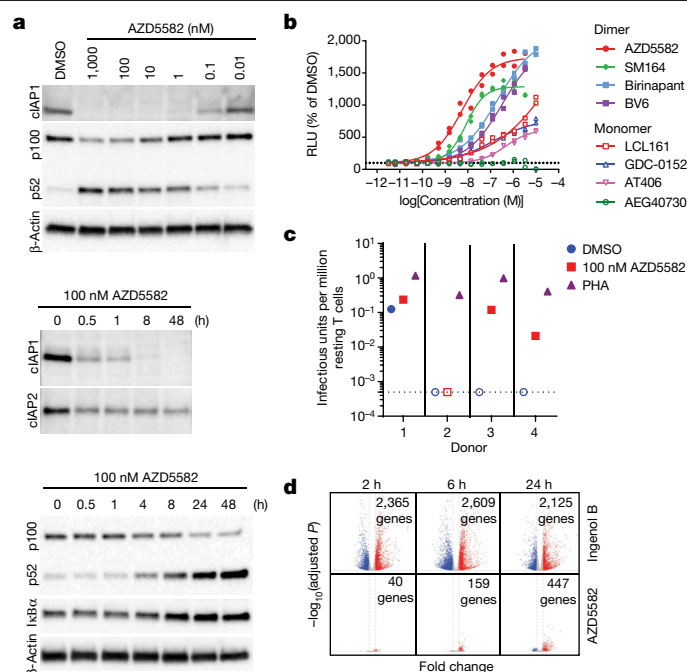
Latently infected cells carrying an integrated replication-competent provirus that contribute to viral rebound after the interruption of ART (termed the HIV reservoir) are not detected and eliminated by the immune system or current therapeutics. Therefore, the HIV reservoir has been targeted by approaches to reverse latency and induce viral antigen production (that is, ‘HIV reactivation’)<sup>2–9</sup>, which renders infected cells susceptible to virus-induced cell death or clearance by the immune system. Previous approaches to HIV reactivation have

been modestly effective and have not demonstrated reactivation of HIV in resting CD4<sup>+</sup> T cells in tissues<sup>2–9</sup>.

## HIV induction in vitro by SMAC mimetics

The lack of specificity of molecules that activate the NF- $\kappa$ B pathway as latency-reversal agents (LRAs) often leads to toxicities that prevent clinical implementation<sup>10</sup>. We tested the induction of HIV and

<sup>1</sup>International Center for the Advancement of Translational Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>2</sup>Division of Infectious Diseases, Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>3</sup>Center for AIDS Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>4</sup>Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA. <sup>5</sup>UNC HIV Cure Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>6</sup>Qura Therapeutics, Chapel Hill, NC, USA. <sup>7</sup>HIV Drug Discovery, ViiV Healthcare, Research Triangle Park, NC, USA. <sup>8</sup>Yerkes National Primate Research Center, Emory University, Atlanta, GA, USA. <sup>9</sup>Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>10</sup>Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>11</sup>GlaxoSmithKline Research and Development, Collegeville, PA, USA. <sup>12</sup>Department of Microbiology and Immunology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>13</sup>AIDS and Cancer Virus Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. <sup>14</sup>Department of Infectious Diseases, Guangzhou Eighth People's Hospital, Guangzhou Medical University, Guangzhou, China. <sup>15</sup>Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA. <sup>16</sup>Chronic Viral Infection Service, McGill University Health Centre, Montreal, Quebec, Canada. <sup>17</sup>Division of Hematology, McGill University Health Centre, Montreal, Quebec, Canada. <sup>18</sup>Department of Epidemiology, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. <sup>19</sup>Emory + Children's Center for Childhood Infections and Vaccines, Atlanta, GA, USA. <sup>20</sup>These authors contributed equally: Christopher C. Nixon, Maud Mavigner. <sup>21</sup>These authors jointly supervised this work: Richard M. Dunham, Ann Chahroudi, J. Victor Garcia. \*e-mail: richard.m.dunham@viivhealthcare.com; ann.m.chahroudi@emory.edu; victor\_garcia@med.unc.edu



**Fig. 1 | Efficient in vitro AZD5582 target engagement and induction of HIV transcription.** **a**, Total CD4<sup>+</sup> T cells were treated with a broad range of concentrations (10 pM–1  $\mu$ M) of AZD5582 overnight, and cell lysates were analysed by immunoblot, probing for cIAP1 and p100/p52 as indicated (top; representative of 10 experiments). Immunoblot analysis of isolated total CD4<sup>+</sup> T cell lysates after treatment with 100 nM AZD5582, examining components of the cNF- $\kappa$ B and ncNF- $\kappa$ B pathways over a 48-h time course after treatment (middle and bottom; representative of three and four experiments, respectively). **b**, DMSO-normalized reporter signal induced by a dose titration of a panel of mono- and bivalent SMAC mimetics in a Jurkat luciferase reporter model of HIV-1 latency with 48 h exposure. Symbols represent technical replicates from a single run and are representative of three independent experiments. Lines represent a four-parameter logistic-regression model fit. **c**, Infectious units per million resting CD4<sup>+</sup> T cells induced by DMSO or 100 nM AZD5582 were determined in a limiting dilution quantitative viral outgrowth assay. PHA, phytohaemagglutinin. **d**, Volcano plots summarizing mean up- and downregulated genes at 2, 6 and 24 h after treatment with ingenol B or AZD5582 compared with treatment with DMSO alone. The mean  $\log_2$ -transformed fold change is shown on the x axis and  $\log_{10}$ -adjusted *P* values (two-sided Wald test) are shown on the y axis. Dashed lines represent thresholds of  $\log_2$ -transformed fold change of 1 and adjusted *P* < 0.05. The data shown represent the mean fold change across four donors and one experiment. For gel source data, see Supplementary Fig. 1.

SIV transcription in latently infected cells by the non-canonical (nc) NF- $\kappa$ B pathway. This pathway activates a limited number of cellular genes and a more-gradual but persistent activation of NF- $\kappa$ B-driven transcription than the canonical (c)NF- $\kappa$ B pathway<sup>11</sup>. Mimetics of the second mitochondrial-derived activator of caspases (SMAC) activate the ncNF- $\kappa$ B pathway by inhibiting the cellular inhibitor of apoptosis protein 1 (cIAP1) and cIAP2. cIAP1 continually represses the ncNF- $\kappa$ B pathway by constitutively degrading the NF- $\kappa$ B-inducing kinase, thus preventing processing of p100 into p52<sup>12</sup>; this repression can be relieved in CD4<sup>+</sup> T cells by in vitro treatment with the SMAC mimetic AZD5582 (Fig. 1a and Extended Data Fig. 1). Compared with other SMAC mimetics, AZD5582 had a superior capacity to reverse HIV latency in vitro<sup>13</sup> (Fig. 1b). AZD5582 also induced replication-competent HIV expression in resting CD4<sup>+</sup> T cells from ART-suppressed HIV-infected donors (Fig. 1c). AZD5582 induced five- to tenfold fewer genes than the protein kinase C agonist ingenol B (Fig. 1d), a cNF- $\kappa$ B pathway inducer and activator of several transcription factors. By specifically targeting the

ncNF- $\kappa$ B signalling pathway, AZD5582 has limited pleiotropic effects, which may translate to fewer off-target effects<sup>14</sup>.

## Latency reversal in BLT humanized mice

BLT mice were infected with HIV-1<sub>JR-CSF</sub> (Supplementary Table 1) and suppressed with ART<sup>15–18</sup> (Fig. 2a, b). Mice then received a single intraperitoneal injection of 3 mg kg<sup>−1</sup> AZD5582 or vehicle. No changes in plasma HIV RNA levels were detected in vehicle-control-treated mice at 24 or 48 h nor in AZD5582-treated mice 24 h after AZD5582 administration (Fig. 2c). However, 48 h after AZD5582 treatment increased HIV RNA expression was detected in the plasma of 3 out of 6 (50%) and 3 out of 4 (75%) mice in two independent experiments (Fig. 2c). These data demonstrate that a single dose of AZD5582 can induce HIV production, resulting in significant viraemia (up to 1,574 HIV RNA copies per ml plasma) in ART-treated BLT mice (Supplementary Table 2).

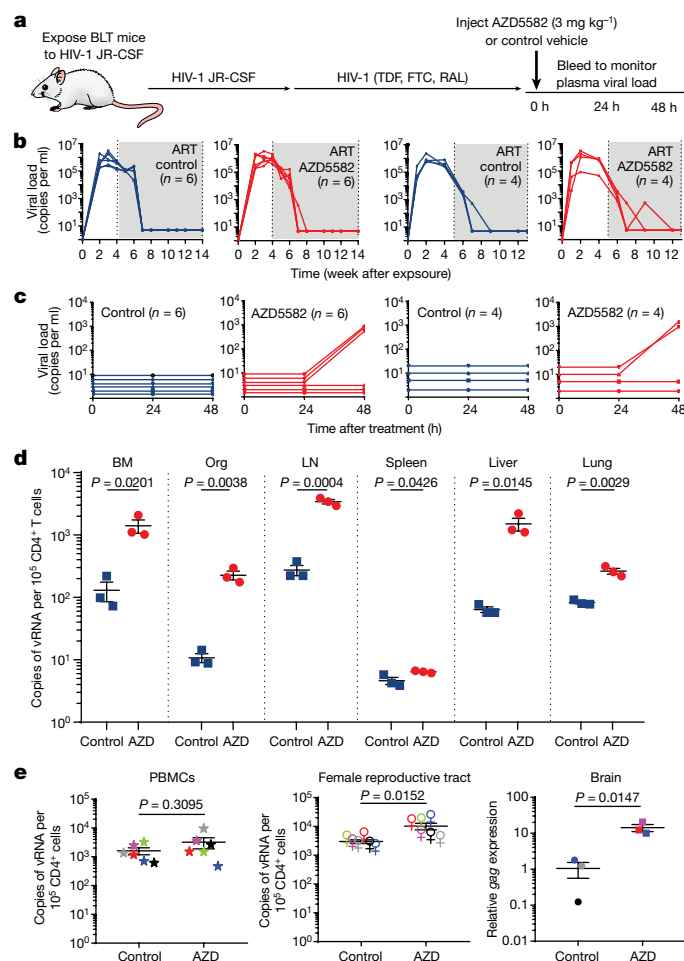
The hallmark of HIV persistence in humans is the presence of inducible HIV in resting CD4<sup>+</sup> T cells. Therefore, resting cells from primary (bone marrow and thymic organoid), secondary (lymph node and spleen) and effector (liver and lung) immune tissues were isolated from HIV-infected ART-suppressed BLT mice 48 h after treatment with vehicle control or AZD5582<sup>17,19,20</sup>. The levels of HIV RNA in resting CD4<sup>+</sup> T cells from AZD5582-treated mice were 11-fold (bone marrow, *P* = 0.0201), 21-fold (thymic organoid, *P* = 0.0038), 12-fold (lymph node, *P* = 0.0004), 1.4-fold (spleen, *P* = 0.0426), 24-fold (liver, *P* = 0.0145) and 3.2-fold (lung, *P* = 0.0029) higher than controls (Fig. 2d). These results were confirmed in a second independent experiment (Extended Data Fig. 2). No notable differences in cell-associated HIV DNA were noted between mice treated with vehicle control or AZD5582 (Supplementary Table 3). These results demonstrate that AZD5582 induces systemic HIV RNA production in resting CD4<sup>+</sup> T cells, indicative of latency reversal in this important cellular source of persistent HIV infection. We also isolated cells from the peripheral blood, female reproductive tract and brain of HIV-infected ART-suppressed BLT mice 48 h after treatment with AZD5582 or vehicle control. As too few CD4<sup>+</sup> T cells were available for cell sorting from these compartments, RNA was extracted from total cells and analysed for the presence of HIV RNA in each tissue. The levels of HIV RNA were significantly higher in the female reproductive tract (3.4-fold, *P* = 0.0152) and brain (8.7-fold, *P* = 0.0147) of AZD5582-treated mice compared with vehicle controls, but not in the blood (*P* = 0.3095) (Fig. 2e). Together, these results show that AZD5582 treatment induces systemic HIV RNA production in BLT mice.

## Pharmacodynamics and safety in BLT mice

Target engagement after treatment with AZD5582 was confirmed ex vivo by the degradation of cIAP1 (proximal) and p100 (distal), targets of SMAC in the ncNF- $\kappa$ B pathway (Extended Data Fig. 3a). In vivo target engagement was demonstrated in resting CD4<sup>+</sup> T cells isolated from thymus, spleen, lymph nodes, liver, lung and bone marrow of BLT mice treated with a single dose of 3 mg kg<sup>−1</sup> AZD5582 or vehicle control (Extended Data Fig. 3b). These results were confirmed by immunohistochemical analysis of the thymic organoid of HIV-infected ART-suppressed BLT mice. This analysis showed a marked reduction in cIAP1 expression in the thymic organoid of AZD5582-treated mice (Extended Data Fig. 3c).

To study off-target or immune-mediated toxicities of AZD5582, we measured serum chemistry, T cell activation and a panel of plasma cytokines after in vivo treatment of immunocompetent BALB/c mice. AZD5582 administration resulted in mild and transient increases in alanine aminotransferase and aspartate aminotransferase that resolved a few days after treatment. No other changes in serum chemistries were noted (Supplementary Table 4). In addition, no differences were noted in the levels of activated (CD38<sup>+</sup>HLA-DR<sup>+</sup>) CD4<sup>+</sup> or CD8<sup>+</sup> T cells in BLT mice treated with AZD5582 or vehicle control (Supplementary Table 5) or in plasma levels of 41 human cytokines and chemokines that serve as indicators of systemic





**Fig. 2 | AZD5582 induces HIV RNA expression in resting CD4<sup>+</sup> T cells from tissues of HIV-infected ART-suppressed BLT mice.** **a**, BLT mice were infected with HIV-1<sub>JR-CSF</sub>. After 10 weeks of ART treatment, mice received vehicle control or AZD5582. FTC, emtricitabine; RAL, raltegravir; TDF, tenofovir disoproxil fumarate. **b**, HIV RNA copies per ml<sup>-1</sup> of plasma of HIV-infected ART-treated BLT mice before treatment with vehicle control (left; blue lines) or AZD5582 (right; red lines). Two independent experiments were performed (left,  $n = 6$  mice per group; right,  $n = 4$  mice per group). Grey shading, period of ART administration. **c**, Plasma HIV RNA levels in HIV-infected ART-suppressed mice from **b** treated with vehicle control or AZD5582. **d**, HIV viral RNA (vRNA) levels in resting CD4<sup>+</sup> T cells isolated from the bone marrow (BM), thymic organoid (Org), lymph nodes (LN), spleen, liver and lung of control or AZD5582-treated mice (cells pooled from  $n = 6$  mice per group for each tissue) were analysed in triplicate. Data are mean  $\pm$  s.e.m. Statistical significance was determined using a two-sided Student's *t*-test. **e**, Cell-associated HIV RNA copies in the blood ( $n = 6$ ), female reproductive tract ( $n = 6$ ) and brain ( $n = 3$ ). PBMCs, peripheral-blood mononuclear cells. Statistical significance was determined using a two-sided Mann-Whitney test (peripheral-blood mononuclear cells and female reproductive tract) or Student's *t*-test (brain). Colours indicate samples from the same mice. Data are mean  $\pm$  s.e.m.

activation and inflammation<sup>21</sup> (Supplementary Table 6). Together, these results demonstrate that AZD5582 does not cause generalized toxicity or activation of the immune system in the BLT model.

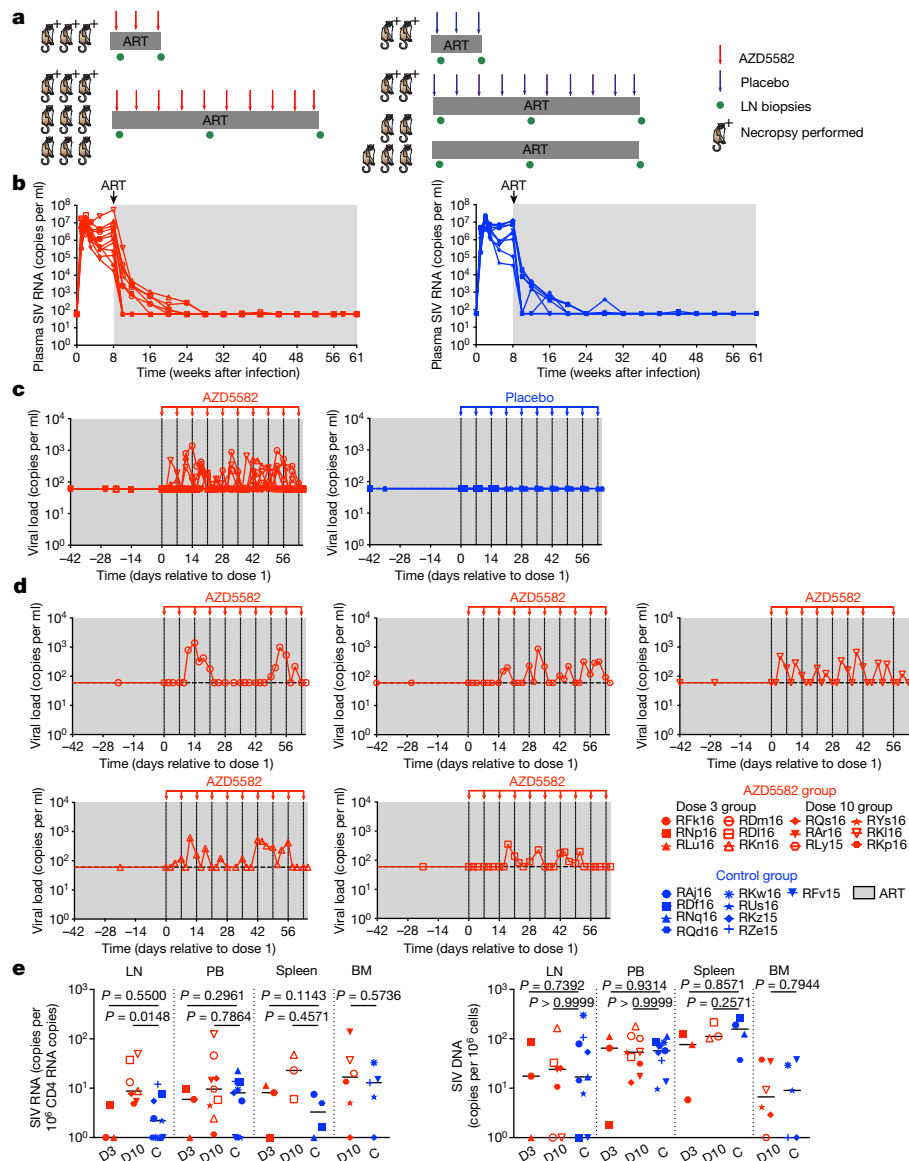
### Latency reversal in rhesus macaques

We next evaluated the latency-reversal activity of AZD5582 in 21 MamuB\*08<sup>-</sup> and MamuB\*17<sup>-</sup> rhesus macaques infected with SIV<sub>mac239</sub> and treated with a potent ART regimen comprising tenofovir disoproxil fumarate, emtricitabine and dolutegravir initiated 8 weeks after

infection (Fig. 3a and Supplementary Table 7)<sup>22,23</sup>. Suppression of SIV viraemia below 60 copies per ml (standard assay limit of detection) was achieved in all macaques in 2–20 weeks and ART was continued for 55–67 weeks before further treatment (Fig. 3b). On the basis of pharmacokinetic and pharmacodynamic data from uninfected macaques (Extended Data Fig. 4a) as well as protocols for SMAC mimetics used in oncology, intravenous infusions of 0.1 mg kg<sup>-1</sup> AZD5582 were administered weekly to 12 SIV-infected ART-suppressed rhesus macaques for 3 or 10 weeks (Fig. 3a). Nine SIV-infected ART-suppressed rhesus macaques served as controls (Fig. 3a). Plasma concentrations of AZD5582 measured after the first, third, sixth and tenth dose showed that drug exposures in SIV-infected ART-suppressed macaques were consistent across the treatment period and comparable to those observed in uninfected rhesus macaques (Extended Data Fig. 4b).

Latency reversal, defined as on-ART viraemia increasing from less than 60 copies per ml of plasma to more than 60 copies per ml of plasma after AZD5582 treatment, was observed as early as 96 h after the first dose and reached levels as high as 1,390 copies per ml in SIV-infected rhesus macaques (Fig. 3c, d). On-ART viraemia >60 copies per ml of plasma was observed in 5 out of 12 rhesus macaques (42%), corresponding to 5 out of 9 rhesus macaques (55%) that received 10 doses of AZD5582 (Fig. 3c, d). Multiple instances of sustained viraemia >60 copies per ml between AZD5582 doses were observed. Out of 140 viral load measurements performed on the 5 macaques that exhibited on-ART viraemia during AZD5582 treatment, 64 were >60 copies per ml (46%); in the macaque with the greatest frequency of reactivation, this proportion was 15 out of 28 (53%). Longitudinal examination of plasma virus by single-genome sequencing analysis of the SIV<sub>mac239</sub> *env* gene in all rhesus macaques that experienced AZD5582-induced on-ART viraemia was performed at four selected time points: 2 weeks after infection (near peak viraemia), 8 weeks after infection (immediately before ART initiation), and at 2 time points separated by 26–42 days during AZD5582 treatment. Phylogenetic analyses showed several patterns of virus reactivation (Extended Data Fig. 5). In two rhesus macaques (RD116 and RKn16), most of the reactivated virus sequences were phylogenetically closer to sequences at eight weeks after infection rather than peak viraemia and were unique, indicating that the variants produced during AZD5582 treatment originated from multiple cells that were seeded at the time of ART initiation<sup>24</sup>. In two other rhesus macaques (RK116 and RDm16), a large fraction of the viruses produced during AZD5582 treatment showed identical sequences, suggesting that latency reversal occurred from a single cell or a clonally expanded population of infected cells. These clones clustered with both peak and pre-ART time points and were accompanied by additional unique sequences. In one rhesus macaque (RLy15), a single virus sequence was amplified at each time point during AZD5582 treatment and these were both phylogenetically similar to sequences found before ART treatment. Taken together, these results indicate that AZD5582 induced virus reactivation from a diverse population of cells, some of which may be clonally expanded<sup>25,26</sup>.

We quantified cell-associated SIV RNA and SIV DNA in resting CD4<sup>+</sup> T cells sorted from SIV-infected ART-suppressed rhesus macaques treated or not with AZD5582. Cell-associated SIV RNA levels in resting CD4<sup>+</sup> T cells isolated from lymph nodes were significantly higher in macaques who received ten doses of AZD5582 compared with controls ( $P = 0.0148$ ) (Fig. 3e). A similar trend was observed in resting CD4<sup>+</sup> T cells isolated from the spleens of a subgroup of six macaques that were euthanized. Levels of cell-associated SIV DNA in resting CD4<sup>+</sup> T cells were similar in each compartment across groups (Fig. 3e). To further understand whether latency reversal induced by AZD5582 resulted in a perturbation of the overall level of infected CD4<sup>+</sup> T cells, we performed longitudinal measurements of cell-associated SIV DNA in total (rather than resting) CD4<sup>+</sup> T cells isolated from lymph nodes and blood as well as quantitative viral outgrowth assays using CD4<sup>+</sup> T cells from lymph nodes and spleen at the end of the treatment period (Extended



**Fig. 3 | AZD5582 induces SIV RNA expression in the plasma and lymph nodes of ART-suppressed SIV-infected rhesus macaques.** **a**, Experimental design during AZD5582 treatment phase. Three rhesus macaques received three doses of AZD5582 and were euthanized 48 h after the last dose. Nine rhesus macaques received 10 doses of AZD5582 and 3 were euthanized 48 h after the last dose. The first dose of AZD5582 was administered after 55–67 weeks of ART (3-dose group, 55 weeks; 10-dose group, 55–67 weeks). Four control rhesus macaques received a weekly placebo infusion; 2 were euthanized 48 h after 3 infusions and 2 were euthanized 48 h after 10 infusions. Five control macaques received ART only. **b**, Plasma SIV RNA levels in the 21 SIV-infected rhesus macaques before treatment with AZD5582 (left,  $n = 12$ ) and equivalent time period for controls (right,  $n = 9$ ). **b–d**, Grey shading represents the period of ART administration. **c**, Plasma SIV RNA levels in ART-suppressed SIV-infected rhesus macaques during AZD5582 treatment (left,  $n = 12$ ) and the equivalent

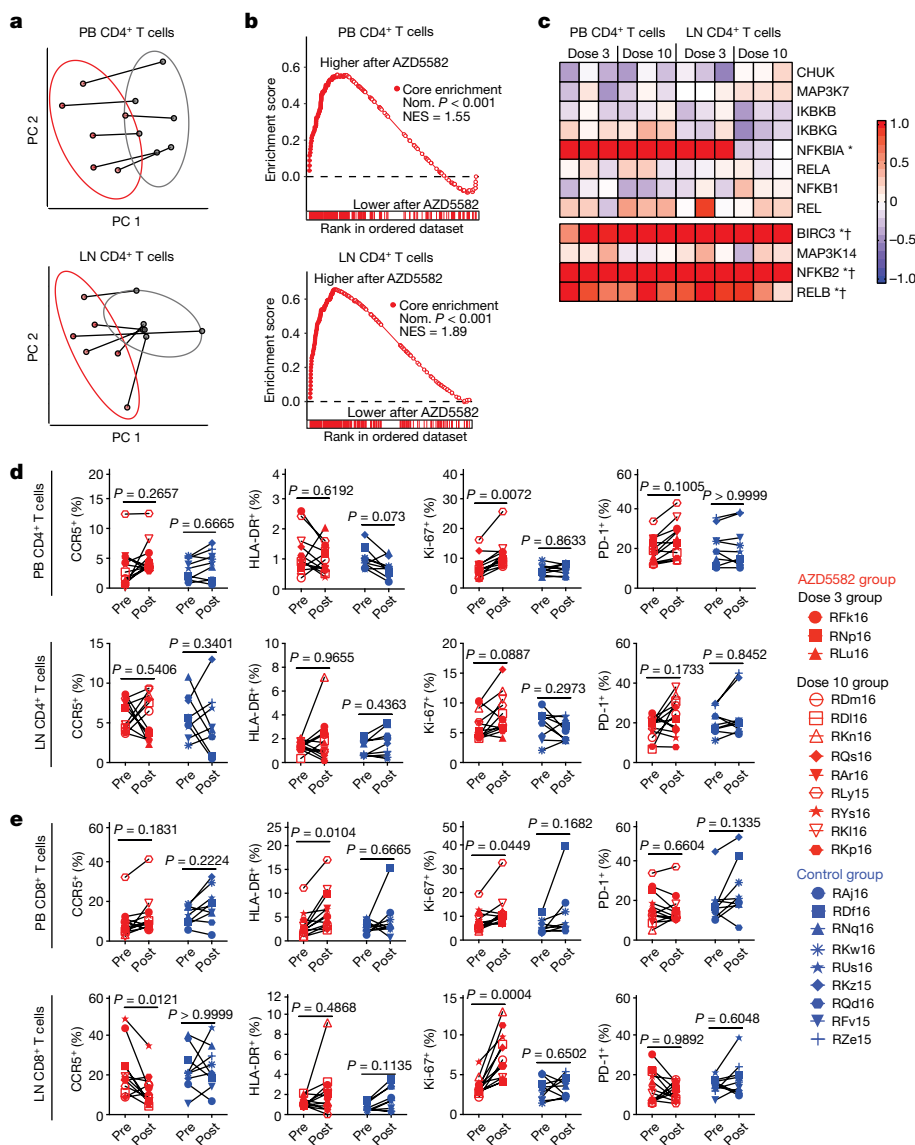
time period for controls (right,  $n = 9$ ). **d**, Individual representation of plasma SIV RNA levels in the five rhesus macaques that experienced on-ART viraemia during AZD5582 treatment. **e**, Cell-associated SIV RNA (left) and SIV DNA (right) levels in resting CD4<sup>+</sup> T cells isolated from lymph nodes, peripheral blood (PB), spleen and bone marrow of AZD5582-treated (red) and control (blue) ART-suppressed SIV-infected rhesus macaques. Resting CD4<sup>+</sup> T cells were analysed from AZD5582-treated rhesus macaques 48 h after receiving 3 doses (D3; lymph nodes, peripheral blood and spleen,  $n = 3$ ) or 10 doses (D10; lymph nodes,  $n = 7$ ; peripheral blood,  $n = 9$ ; spleen,  $n = 3$ ; bone marrow,  $n = 6$ ) of AZD5582. Resting CD4<sup>+</sup> T cells were analysed from control rhesus macaques (C; lymph nodes and peripheral blood,  $n = 9$ ; spleen,  $n = 4$ ; bone marrow,  $n = 5$ ) at equivalent time points. Open symbols indicate AZD5582-treated rhesus macaques with on-ART viraemia. Statistical significance was determined with a two-sided Mann–Whitney  $U$ -test. Horizontal lines represent the median.

Data Fig. 6a–c). Despite the high level of virus reactivation induced by AZD5582, these experiments did not reveal a consistent reduction in the total or replication-competent SIV reservoir compared with controls.

### Pharmacodynamics in rhesus macaques

Pharmacological target engagement of the  $\text{mNF-}\kappa\text{B}$  pathway was confirmed by western blot analysis of the degradation of p100 to p52

in lymph-node mononuclear cells after in vivo exposure to AZD5582 (Extended Data Fig. 4c) and in splenocytes treated ex vivo with AZD5582 (Extended Data Fig. 4d–g). Transcriptomic profiling of CD4<sup>+</sup> T cells from the peripheral blood and lymph nodes isolated from AZD5582-treated rhesus macaques showed a distinct effect of AZD5582 on gene expression based on principal component (Fig. 4a) and DAVID pathway (Extended Data Fig. 7a) analyses. Enrichment of NF- $\kappa\text{B}$  targets after AZD5582 treatment was demonstrated by gene-set enrichment analysis



**Fig. 4 | AZD5582 specifically activates the ncNF-κB pathway in SIV-infected ART-suppressed rhesus macaques without generalized T cell activation.**

**a–c**, Gene expression in CD4<sup>+</sup> T cells from the peripheral blood and lymph nodes of SIV-infected ART-suppressed rhesus macaques before and after treatment with AZD5582 ( $n = 6$  for both lymph nodes and peripheral blood; for each,  $n = 3$  for 3 doses of AZD5582 and  $n = 3$  for 10 doses of AZD5582). **a**, Principal component (PC) analyses of the transcriptomes of CD4<sup>+</sup> T cells from the peripheral blood (top) and lymph nodes (bottom), before (grey) and after (red) treatment with AZD5582. Ellipses, two standard deviations. **b**, GSEA plots of NF-κB-induced genes in CD4<sup>+</sup> T cells from the peripheral blood (top) and lymph nodes (bottom). Gene set: ‘hallmark TNF signalling via NF-κB’ (MSigDB). NES, normalized enrichment score. **c**, Heat map of cNF-κB (top) and ncNF-κB (bottom) pathway gene expression. Colour scale, log<sub>2</sub>-transformed fold changes after AZD5582 treatment compared with before treatment. Genes that are differentially expressed after treatment with AZD5582 in the peripheral blood (asterisks) or lymph nodes (daggers) CD4<sup>+</sup> T cells are highlighted.

**d, e**, Expression of activation markers in CD4<sup>+</sup> (**d**) and CD8<sup>+</sup> (**e**) T cells in the peripheral blood (top) and lymph nodes (bottom) of SIV-infected ART-suppressed rhesus macaques before treatment with AZD5582 and 48 h after the final dose ( $n = 12$ , red) and control rhesus macaques ( $n = 9$ , blue). Statistical significance was determined using a two-sided Mann–Whitney *U*-test.

(GSEA) (Fig. 4b; a heat map of differentially expressed gene targets of NF-κB is shown in Extended Data Fig. 7b). The hallmark ncNF-κB signalling genes *NFKB2* and *RELB* were significantly upregulated in CD4<sup>+</sup> T cells from the peripheral blood and lymph nodes after AZD5582 treatment, whereas cNF-κB signalling molecules were mostly unaffected except for the inhibitor *NFKBIA* (Fig. 4c). The baculoviral IAP repeat-containing 3 (*BIRC3*) gene, which encodes cIAP2 and regulates both the cNF-κB and ncNF-κB pathways, was also significantly upregulated in CD4<sup>+</sup> T cells from the peripheral blood and lymph nodes (Fig. 4c). Notably, activation of ncNF-κB signalling genes was evident in rhesus macaques with and without on-ART viraemia of more than 60 copies per ml (Fig. 4c and Extended Data Fig. 7c) and GSEA showed similar changes in overall gene expression when AZD5582-treated rhesus macaques were grouped according to the presence or absence of on-ART viraemia of more than 60 copies per ml (Supplementary Table 8), suggesting that the lack of a virological response measured by the standard viral load assay was not due to compromised target engagement.

We next used an ultrasensitive assay to measure SIV RNA in plasma, comparing 2–3 baseline time points during ART in the 4 weeks before AZD5582 treatment (when plasma SIV RNA was less than 60 copies per ml in all rhesus macaques using the standard viral load assay) and 3–4 time points during AZD5582 treatment (Extended Data Fig. 8a). We found significantly higher plasma SIV RNA values during the period

of AZD5582 treatment compared with pre-treatment (Extended Data Fig. 8b,  $P = 0.008$ ), with 8 out of 12 rhesus macaques demonstrating  $\geq 2$  SIV RNA measurements above the median of their baseline values. This group of eight rhesus macaques with evidence of latency reversal induced by AZD5582 could be segregated from the four rhesus macaques with unchanged viral loads based on two parameters: higher levels of plasma SIV RNA immediately before ART initiation and SIV DNA in peripheral CD4<sup>+</sup> T cells before AZD5582 treatment ( $P = 0.004$  for each) (Extended Data Fig. 8c).

### Safety and immune effects in rhesus macaques

The potential for AZD5582 toxicity in SIV-infected ART-suppressed rhesus macaques was examined and transient increases in liver enzymes (aspartate aminotransferase and  $\gamma$ -glutamyltransferase) were observed (Extended Data Fig. 9a). Total white-blood cell counts were decreased in all rhesus macaques 48 h after the first dose of AZD5582 but returned to normal levels when measured before the second dose and over time (Extended Data Fig. 9b). After the seventh and eighth doses, one macaque experienced a reaction characterized by fever, emesis, fatigue and lack of appetite, and had elevated liver enzyme and creatinine levels and bandaemia as shown by laboratory examination. All abnormalities resolved within a two-week period; however, this rhesus macaque did

not receive further doses. Notably, weights did not significantly fluctuate over the course of the treatment period (Extended Data Fig. 9c). In summary, 97 doses of AZD5582 were administered to 12 SIV-infected ART-suppressed rhesus macaques, of which 95 were well-tolerated and 2 resulted in a mild adverse reaction.

Markers of T cell activation (both CD4<sup>+</sup> and CD8<sup>+</sup>) were assessed in AZD5582-treated rhesus macaques and controls, and expression of CCR5, HLA-DR and PD-1 on CD4<sup>+</sup> T cells from the blood and lymph nodes was similar before and after treatment with AZD5582 (Fig. 4d, e). Levels of intracellular Ki-67 expression were increased in CD4<sup>+</sup> T cells in the blood and CD8<sup>+</sup> T cells in the blood and lymph nodes after AZD5582 treatment (Fig. 4d, e). CD8<sup>+</sup> T cell expression of HLA-DR was higher in the blood after exposure to AZD5582 (Fig. 4e). These results indicate that AZD5582 does not induce global CD4<sup>+</sup> T cell activation in rhesus macaques but the combined pharmacological and virological effects may have a stimulatory effect on CD8<sup>+</sup> T cells. Longitudinal analyses of CD4<sup>+</sup> T cell counts and frequencies, CD4<sup>+</sup> T cell viability, CD4<sup>+</sup> T cell subset frequencies and their Ki-67 expression are shown in Extended Data Fig. 10a–d. The increased Ki-67 observed within CD4<sup>+</sup> T cells after AZD5582 treatment may provide a note of caution as the proliferation of latently infected memory CD4<sup>+</sup> T cells is hypothesized to contribute to the maintenance of the viral reservoir over time<sup>25,27</sup>; however, we did not find an increase in infected cells after AZD5582 treatment (Fig. 3e and Extended Data Fig. 6).

We next determined whether AZD5582 would impair SIV-specific T cell responses in SIV-infected ART-suppressed rhesus macaques, a potential adverse outcome of treatment that has been suggested for other LRAs<sup>28</sup>. However, the frequency of SIV Gag- or Env-specific CD8<sup>+</sup> T cells measured by IFN $\gamma$  ELISPOT did not decrease after AZD5582 treatment (Extended Data Fig. 10e), and CD8<sup>+</sup> T cell polyfunctionality and proliferative responses were largely unaffected by AZD5582 treatment ex vivo, with the exception of IL-2<sup>+</sup>IFN $\gamma$ <sup>+</sup>TNF<sup>+</sup> triple-positive cells, which were slightly reduced ( $P = 0.0476$ ) (Extended Data Fig. 10f, g). Furthermore, longitudinal assessment of plasma levels of inflammatory cytokines and chemokines did not reveal any marked changes induced by AZD5582 (Extended Data Fig. 10h). Overall, our work in the rhesus macaque model indicates that AZD5582 treatment is safe in most macaques and can induce appreciable increases in plasma SIV RNA as well as SIV RNA expression in resting CD4<sup>+</sup> T cells from the lymph nodes during ART.

## Discussion

Eradication of HIV infection after prolonged viral suppression is the focus of intense research and latency reversal has been a cornerstone of this effort. LRAs have been widely recognized as important (but previously mostly theoretical) tools to induce HIV expression in resting CD4<sup>+</sup> T cells in humans. Future clinical applications of HIV cure strategies must be relevant to the majority of people living with HIV for whom the treatments of the Berlin and London patients<sup>29,30</sup> (whose non-HIV life-threatening haematological malignancies warranted aggressive, toxic therapies) pose an unacceptable level of risk. Therefore, LRAs must be identified that are highly effective but have minimal side effects. Here, we used two different but highly complementary animal models<sup>31</sup> to show that treatment with AZD5582 had minimal and transient side effects but resulted in significant increases in HIV and SIV RNA levels both in the plasma and in resting CD4<sup>+</sup> T cells isolated from all analysed tissues from BLT mice and from the lymph nodes of macaques. Our results provide in vivo evidence of systemic HIV and SIV latency reversal from resting CD4<sup>+</sup> T cells. The concordance between the results obtained in two fundamentally different animal models highlights the robust and reproducible nature of the effect of AZD5582 on HIV and SIV reservoirs. The fact that there is little toxicity associated with the use of AZD5582 strongly suggests that activators of the  $\text{mTORC1}$  pathway may be well-suited for HIV eradication approaches in humans.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1951-3>.

1. Finzi, D. et al. Latent infection of CD4<sup>+</sup> T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **5**, 512–517 (1999).
2. Archin, N. M. et al. Interval dosing with the HDAC inhibitor vorinostat effectively reverses HIV latency. *J. Clin. Invest.* **127**, 3126–3135 (2017).
3. Archin, N. M. et al. Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487**, 482–485 (2012).
4. Elliott, J. H. et al. Activation of HIV transcription with short-course vorinostat in HIV-infected patients on suppressive antiretroviral therapy. *PLoS Pathog.* **10**, e1004473 (2014).
5. Gutiérrez, C. et al. Bryostatins for latent virus reactivation in HIV-infected patients on antiretroviral therapy. *AIDS* **30**, 1385–1392 (2016).
6. Kulkosky, J. et al. Intensification and stimulation therapy for human immunodeficiency virus type 1 reservoirs in infected persons receiving virally suppressive highly active antiretroviral therapy. *J. Infect. Dis.* **186**, 1403–1411 (2002).
7. Prins, J. M. et al. Immuno-activation with anti-CD3 and recombinant human IL-2 in HIV-1-infected patients on potent antiretroviral therapy. *AIDS* **13**, 2405–2410 (1999).
8. Rasmussen, T. A. et al. Panobinostat, a histone deacetylase inhibitor, for latent-virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: a phase 1/2, single group, clinical trial. *Lancet HIV* **1**, e13–e21 (2014).
9. Søgaard, O. S. et al. The depsipeptide romidepsin reverses HIV-1 latency in vivo. *PLoS Pathog.* **11**, e1005142 (2015).
10. Ke, R., Conway, J. M., Margolis, D. M. & Perelson, A. S. Determinants of the efficacy of HIV latency-reversing agents and implications for drug and treatment design. *JCI Insight* **3**, e123052 (2018).
11. Sun, S. C. The noncanonical NF- $\kappa$ B pathway. *Immunol. Rev.* **246**, 125–140 (2012).
12. Fulda, S. Molecular pathways: targeting death receptors and Smac mimetics. *Clin. Cancer Res.* **20**, 3915–3920 (2014).
13. Pache, L. et al. BIRC2/cIAP1 is a negative regulator of HIV-1 transcription and can be targeted by Smac mimetics to promote reversal of viral latency. *Cell Host Microbe* **18**, 345–353 (2015).
14. Hennessy, E. J. et al. Discovery of a novel class of dimeric Smac mimetics as potent IAP antagonists resulting in a clinical candidate for the treatment of cancer (AZD5582). *J. Med. Chem.* **56**, 9897–9919 (2013).
15. Honeycutt, J. B. et al. T cells establish and maintain CNS viral infection in HIV-infected humanized mice. *J. Clin. Invest.* **128**, 2862–2876 (2018).
16. Kessing, C. F. et al. In vivo suppression of HIV rebound by didehydro-cortistatin A, a “block-and-lock” strategy for HIV-1 treatment. *Cell Reports* **21**, 600–611 (2017).
17. Tsai, P. et al. In vivo analysis of the effect of panobinostat on cell-associated HIV RNA and DNA levels and latent HIV infection. *Retrovirology* **13**, 36 (2016).
18. Melkus, M. W. et al. Humanized mice mount specific adaptive and innate immune responses to EBV and TSST-1. *Nat. Med.* **12**, 1316–1322 (2006).
19. Choudhary, S. K. et al. Latent HIV-1 infection of resting CD4<sup>+</sup> T cells in the humanized Rag2<sup>-/-</sup>  $\gamma$ - $\gamma$  mouse. *J. Virol.* **86**, 114–120 (2012).
20. Denton, P. W. et al. Generation of HIV latency in humanized BLT mice. *J. Virol.* **86**, 630–634 (2012).
21. Wahl, A. et al. Precision mouse models with expanded tropism for human pathogens. *Nat. Biotechnol.* **37**, 1163–1173 (2019).
22. Mavigner, M. et al. Simian immunodeficiency virus persistence in cellular and anatomic reservoirs in antiretroviral therapy-suppressed infant rhesus macaques. *J. Virol.* **92**, e00562-18 (2018).
23. Mavigner, M. et al. Pharmacological modulation of the Wnt/ $\beta$ -catenin pathway inhibits proliferation and promotes differentiation of long-lived memory CD4<sup>+</sup> T cells in antiretroviral therapy-suppressed simian immunodeficiency virus-infected macaques. *J. Virol.* **94**, e01094-19 (2019).
24. Abrahams, M. R. et al. The replication-competent HIV-1 latent reservoir is primarily established near the time of therapy initiation. *Sci. Transl. Med.* **11**, eaaw5589 (2019).
25. Anderson, E. M. & Maldarelli, F. The role of integration and clonal expansion in HIV infection: live long and prosper. *Retrovirology* **15**, 71 (2018).
26. Ferris, A. L. et al. Clonal expansion of SIV-infected cells in macaques on antiretroviral therapy is similar to that of HIV-infected cells in humans. *PLoS Pathog.* **15**, e1007869 (2019).
27. Kuo, H. H. & Lichterfeld, M. Recent progress in understanding HIV reservoirs. *Curr. Opin. HIV AIDS* **13**, 137–142 (2018).
28. Clutton, G. T. & Jones, R. B. Diverse impacts of HIV latency-reversing agents on CD8<sup>+</sup> T-cell function: implications for HIV cure. *Front. Immunol.* **9**, 1452 (2018).
29. Gupta, R. K. et al. HIV-1 remission following CCR5 $\Delta$ 32/ $\Delta$ 32 haematopoietic stem-cell transplantation. *Nature* **568**, 244–248 (2019).
30. Hütter, G. et al. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N. Engl. J. Med.* **360**, 692–698 (2009).
31. Nixon, C. C., Mavigner, M., Silvestri, G. & Garcia, J. V. In vivo models of human immunodeficiency virus persistence and cure strategies. *J. Infect. Dis.* **215**, S142–S151 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020



## Methods

### Experimental design

The purpose of this study was to determine the efficacy of the SMAC-mimetic AZD5582 as an HIV and SIV LRA *in vivo*. To this end, two animal models of HIV-1 infection were used: the HIV-infected BLT humanized mouse model and the SIV-infected rhesus macaque (*Macaca mulatta*) nonhuman primate (NHP) model. In each model, animals were infected with HIV-1 or SIV and viraemia was durably suppressed with ART. AZD5582 was administered to infected ART-suppressed animals that were then assessed for changes associated with the reactivation of the viral reservoir. Mice were maintained under specific-pathogen-free conditions by the Division of Comparative Medicine at the University of North Carolina, Chapel Hill. Mouse experiments were conducted in accordance with NIH guidelines for the housing and care of laboratory animals and in accordance with protocols reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) at the University of North Carolina, Chapel Hill. Healthy rhesus macaques for pharmacokinetic studies were housed at GlaxoSmithKline and all procedures were conducted in accordance with the GlaxoSmithKline policy on the care, welfare and treatment of laboratory animals and were reviewed by the IACUC at GlaxoSmithKline. Rhesus macaques infected with SIV were housed at the Yerkes National Primate Research Center and treated in accordance with Emory University and Yerkes National Primate Research Center IACUC regulations (PROTO201800308). Animal-care facilities are accredited by the US Department of Agriculture and the Association for Assessment and Accreditation of Laboratory Animal Care International.

### Preparation of Jurkat HIV-luciferase cell clones

Cell clones Jurkat-C16 and Jurkat-I15 were prepared by infecting Jurkat cells (Jurkat clone E6-1 cells, American Type Culture Collection TIB-152, authenticated by morphological identification and virus-susceptibility profiles, tested for mycoplasma by the supplier) with a full-length, infectious HIV-1<sub>NL4-3</sub>-based virus engineered to express a luciferase reporter in place of the HIV-1 *nef* gene (NLCH-Luci). The Jurkat-N6 cell clone was generated using the same virus as described above with an additional mouse heat-stable antigen (HSA) reporter located just downstream of the luciferase open-reading frame and separated by a T2A element (NLCH-Luci-HSA). NLCH, provided by R. Swanstrom, is the parent molecular infectious clone used to make the Jurkat clones and is a modification of HIV-1<sub>NL4-3</sub> (GenBank U26942) in which flanking sequences were removed. All viruses were derived by transfection of human embryonic kidney 293 cells (HEK 293T, European Collection of Authenticated Cell Cultures, authenticated by morphological identification, tested for mycoplasma by the supplier) with 1 µg of the HIV-1<sub>NL4-3</sub>-derived infectious molecular plasmid DNA using the FuGENE HD Transfection reagent (Promega) according to the manufacturer's recommendations. Supernatants were collected 48 h after transfection, passed through a 0.2-µm filter, and used to infect wild-type Jurkat cells. After infection, cells expressing high levels of HIV-encoded mouse HSA were removed using biotin-labelled rat anti-mouse CD24 antibody (clone M1/69, BD Biosciences) that was adsorbed to streptavidin-labelled magnetic Dynabeads M-280 (Life Technologies) according to the manufacturer's recommendations. Negatively selected HIV-infected Jurkat cells were then limit-diluted to 0.5 cells per well in 96-well plates, and individual cell clones were expanded for 2–4 weeks in culture in the presence of 500 nM efavirenz. Clones were profiled for baseline reporter level and responsiveness to benchmark LRAs, with C16, I15 and N6 representing the most quiescent but inducible clones that were obtained.

### Cell culture and Jurkat HIV-luciferase assay

Jurkat HIV-luciferase clones were maintained in Roswell Park Memorial Institute (RPMI) 1640 medium (Gibco, Life Technologies) containing

10% (v/v) fetal bovine serum (FBS; SAFC, Sigma-Aldrich), 25 U ml<sup>-1</sup> penicillin and 25 U ml<sup>-1</sup> streptomycin (Gibco, Life Technologies), and were split 1:4 every 3–4 days to maintain a cell density of around 0.3–1 million cells per ml. The Jurkat clones were maintained with the addition of 500 nM efavirenz to the medium. Three Jurkat cell clones (C16, I15 and N6), each of which contained one or two integrated HIV proviruses that expressed the luciferase reporter gene, were added at equal amounts for a total of 5,000 cells per well to 384-well plates containing compound titrations. Dose–response testing was performed on compounds dissolved in dimethyl sulfoxide (DMSO; Fisher Scientific) dispensed in duplicate serial threefold, 14-point titrations using a D300e Digital Droplet Dispenser (Hewlett-Packard) to give final assay concentrations ranging from 10 µM to 2.1 pM in 50 µl of medium with a final concentration of 0.5% DMSO (v/v). Cells and compound were incubated at 37 °C for 48 h, unless otherwise indicated, followed by the addition of 20 µl of Steady-Glo Luciferase (Promega). Luminescence resulting from the induction of virally expressed luciferase was measured using an EnVision 2102 Multilabel Plate Reader (Perkin Elmer). Dose–response relationships were analysed with GraphPad Prism (v.6) using a four-parameter logistic regression model to calculate the concentration of compound that provides the half-maximal response and the maximal percentage activation compared to the vehicle control.

### Immunoblot analyses

For the immunoblot assays, 10 µg of cell lysate was loaded per well into 4–20% Tris-Glycine SDS–PAGE gels. Proteins from the SDS–PAGE gels were transferred to Turbo Midi PVDF Transfer Packs (BioRad) using the 'Mixed MW' protocol for one Midi Format Gel (constant 2.5 A up to 25 V, for 7 min) of the Trans-Blot Turbo Transfer System (BioRad) with pre-made Trans-Blots according to the manufacturer's instructions. After transfer, PVDF membranes were blocked in 5% bovine serum albumin (BSA) in 1× Tris-buffered saline (TBS) (BioRad) with 0.1% Tween-20 for 1 h at room temperature with gentle rocking. Primary antibodies were added and incubated overnight at 4 °C (anti-cIAP1, 1:1,000 (Abcam); anti-p100/p52, 1:1,000 (Cell Signaling Technology); anti-IκBα, 1:1,000 (Cell Signaling Technology); anti-cIAP2, 1:1,000 (Abcam); and anti-actin-HRP conjugate, 1:30,000 (Abcam)). After staining with primary antibodies, the membrane was washed three times with 1× TBS and 0.1% Tween-20, 10 min each wash. After washing, the membrane was incubated in 5% BSA in 1× TBS and 0.1% Tween-20 with the appropriate secondary antibody for 2 h at room temperature. After staining with secondary antibodies, the membrane was washed twice for 10 min with 1× TBS and 0.1% Tween-20 followed by a 10-min wash with 1× TBS. The membrane was then patted dry with filter paper and an image was captured of the undeveloped membrane on a ChemiDoc MP Imaging System using Image Laboratory software (v.6.0.1, BioRad). Sufficient ECL reagent (GE Healthcare) was used to cover the membrane and a series of images was taken with increasing exposure times until the luminescence from the developed membrane saturated the image. The developed membrane was then washed 3 times with 1× TBS for 5 min to remove the residual ECL reagent and then stored at 4 °C in sufficient 1× TBS to submerge the entire membrane. Densitometry of images of the developed membrane was then carried out using Image Laboratory software (v.6.0.1, BioRad). Some membranes were stripped for 1 min with One Minute Plus Western Blot Stripping Buffer (GM Biosciences) and then washed 3 times for 10 min with 1× TBS. The stripped membranes were then blocked in 5% BSA in 1× TBS and 0.1% Tween-20 for 1 h and re-probed overnight with a new primary antibody. To normalize samples for loading a 1:20,000 dilution of β-actin (Abcam) was run on the stripped membranes.

### Target gene RT–qPCR

We treated 2 million normal donor CD4<sup>+</sup> T cells with a range of concentrations of AZD5582. Total RNA was isolated using the RNEasy Mini kit (Qiagen) according to the manufacturer's instructions. The following



TaqMan primer probe sets were sourced from Applied Biosystems: Hs00985031\_g1 (*BIRC3*), Hs00174517\_m1 (*NFKB2*) and Hs02800695 (*HPRT1*). TaqMan-based quantitative PCR with reverse transcription (RT-qPCR; Fast Virus 1-Step Master Mix, Applied Biosystems) was used to amplify host genes of interest and acquire the signal on a QuantStudio 3 Real-Time PCR thermocycler (ThermoFisher). Gene expression was normalized to *HPRT1* and the comparative threshold cycle ( $C_t$ ) method ( $\Delta\Delta C_t$ ) was used for relative quantification of gene expression. Relative quantification was analysed by QuantStudio 3 Real-Time PCR System software (v.1.4.3, ThermoFisher).

### HIV quantitative viral outgrowth

All human peripheral-blood mononuclear cell (PBMC) samples were obtained under a specimen procurement protocol reviewed and approved by the University of North Carolina Biomedical Institutional Review Board and the McGill University Health Centre Ethical Review Board. Informed consent was obtained from all participants. Human PBMCs for quantitative viral outgrowth were obtained using continuous flow leukapheresis. Resting CD4<sup>+</sup> T cells were isolated and virus outgrowth assays were performed as previously described<sup>32,33</sup> with some modifications. In brief, 20–50 × 10<sup>6</sup> highly purified resting CD4<sup>+</sup> T cells were stimulated with phytohaemagglutinin, IL-2 (60 U ml<sup>-1</sup>) and irradiated PBMCs from a seronegative donor, or with 100 nM AZD5582, 335 nM vorinostat or 0.003% DMSO (vehicle control) in limiting dilutions for 24 h. Cultures were washed to remove drugs and CCR5<sup>high</sup>, CD8-depleted, phytohaemagglutinin-stimulated PBMCs from an uninfected donor were added twice to amplify virus outgrowth. Culture supernatants were assayed for HIV p24 expression by ELISA on day 15 and confirmed on day 19. A maximum-likelihood method was used to estimate the frequency of inducible virus and is reported as infectious units per million<sup>34</sup>.

### RNA-sequencing analysis of human cells

Total CD4<sup>+</sup> T cells were isolated from PBMCs of four ART-treated aviraemic patients by negative selection (EasySep Human CD4<sup>+</sup> T Cell Enrichment Kit, StemCell) according to the manufacturer's instructions. Dead cells and other debris were removed using a Dead Cell Removal Kit (Miltenyi Biotec) according to the manufacturer's instructions. Cells from each patient were treated with 0.05% DMSO, 100 nM AZD5582 or 25 nM ingenol B and collected 2 h, 6 h and 24 h after exposure. RNA was isolated from the collected cells using AllPrep DNA/RNA Mini Kit (Qiagen). Then, 200 ng of RNA from each sample was checked for quality using an Agilent Bioanalyzer; RNA-integrity number scores were typically >9.0, suggesting that high-quality RNA was obtained. These total RNA samples were then processed into stranded, mRNA libraries using the KAPA library preparation kit (KAPA BioSystems, F. Hoffmann-La Roche). The concentrations of the final libraries were checked by Qubit (Thermo Fisher) and the fragment size distribution (mean size, 359 bp) was analysed with a BioAnalyzer HS-DNA chip (Agilent). Samples were then sequenced using an Illumina HiSeq 4000 sequencer using a paired-end 50-bp by 50-bp run. Samples were successfully demultiplexed and then quality assurance and quality control was carried out using FASTQC (v.0.11.1) (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>).

Raw reads were mapped to the human genome and transcriptome (GRCh38.p7) using STAR<sup>35</sup> and Salmon (v.0.7.2)<sup>36</sup>. Data were normalized and analysed for changes in gene expression using the DESeq2 package<sup>37</sup> in R. *P* values were adjusted for multiple testing using a false-discovery rate using the Benjamini–Hochberg method<sup>38</sup>. Data were analysed both jointly and within each treatment compared with the vehicle control. Differential expression of outliers was assessed and found to have a non-significant overall effect. Thresholds applied to call a significant response were mean log<sub>2</sub>-transformed fold change >1 and adjusted *P* < 0.05. Graphs and summary tables were built in R using ggplot. Gene-set enrichment was performed using GSEA (v.2.2.3)

and GO analysis (GO PANTHER v.11.1)<sup>39</sup>. Results shown are the mean responses of the four donors tested.

### Generation and maintenance of BLT mice

BLT mice were prepared as previously reported<sup>18,40,41</sup>. In brief, a 1–2-mm piece of human liver tissue was sandwiched between two pieces of autologous thymus tissue (Advanced Bioscience Resources) under the kidney capsule of sublethally irradiated (200 cGy) 12–15-week-old female NOD.Cg-*Prkdc*<sup>scid</sup> *IL2rg*<sup>tm1Wjl</sup>/SzJ (NSG; The Jackson Laboratory) mice. After implantation, mice were transplanted intravenously with CD34<sup>+</sup> haematopoietic stem and progenitor cells isolated from autologous human liver tissue. Human immune cell reconstitution was monitored in the peripheral blood of BLT mice by flow cytometry every 3–4 weeks<sup>21</sup>. For the study that examined the effect of AZD5582 administration on T cell activation, the mean weight of the mice used was 26.88 g and they were approximately 1 year of age at the initiation of the study. For the study that examined the effect of AZD5582 administration on plasma and tissue viraemia during ART suppression, the mean weight of the mice used was 23.14 g and the mice were approximately 7 months old at the initiation of the study. Mice were randomized for assignment to either experimental or control groups (<https://www.random.org/>).

### HIV infection of BLT mice

Stocks of HIV-1<sub>JR-CSF</sub> were prepared as follows. The proviral clone was transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocols. Viral supernatant was collected 48 h after transfection. Viral supernatant was titrated by infecting TZM-bl cells (NIH AIDS Reagent Program, authenticated by morphological identification and virus-susceptibility profiles, tested for mycoplasma by the supplier) at multiple dilutions. Virus-containing medium was removed the next day and replaced with fresh Dulbecco's modified Eagle medium (DMEM; ThermoFisher) plus 10% FBS and the incubation continued for 24 h. The cells were fixed and stained with 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside and blue cells were counted directly to determine infectious particles per ml. Each titre of these viral stocks was performed in triplicate and at least two different titre determinations were performed for each virus stock. Exposure of BLT mice to HIV-1<sub>JR-CSF</sub> was conducted by tail vein injection with 3 × 10<sup>4</sup> tissue culture infectious units of virus. Plasma viral load in peripheral blood of infected mice was monitored longitudinally by RT-qPCR using TaqMan RNA to-C<sub>T</sub> 1-step kit (Applied Biosystems). The sequences of the forward and reverse primers and the TaqMan probe for PCR amplification and detection of HIV *gag* RNA were: 5'-CATGTTTTCAGCATTATCAGAAGGA-3', 5'-TGCTTGATGTCCCCCACT-3' and 5'-FAM-CCACCCCAACAAGATTAAACACCAT-GCTAA-Q-3', respectively. For viral load analysis, 40 μl of plasma was collected and analysed with a sensitivity of 350 copies per ml. All samples were run and analysed on an ABI 7500 Fast Real Time PCR System (Applied Biosystems).

### SIV infection of NHPs

In total, 21 male and female Indian rhesus macaques, 3–6 years of age, with the exclusion of MamuB\*08<sup>+</sup> and MamuB\*17<sup>+</sup> macaques, were included in this study (Supplementary Table 7). Rhesus macaques were infected intravenously with 3 × 10<sup>3</sup> TCID<sub>50</sub> (50% tissue culture infectious dose) of SIV<sub>mac239</sub>. The SIV<sub>mac239</sub> stock was titrated in vitro for viral infectivity by standard end-point titration on CEMx174 cells. The TCID<sub>50</sub> was calculated as previously described<sup>42</sup>. Standard SIV<sub>mac239</sub> plasma viral-load quantification was performed regularly throughout the study and three times per week during the AZD5582 treatment period in the Translational Virology Core Laboratory of the Emory Center for AIDS Research using a standard qPCR assay (limit of detection of 60 copies per ml of plasma) as previously described<sup>43</sup>. Ultra-sensitive SIV<sub>mac239</sub> plasma viral-load quantification (limit of detection of 3 copies per ml of plasma) was performed for 2–3 time points before

# Article

AZD5582 treatment and 3–4 time points during AZD5582 treatment as previously described<sup>44,45</sup>.

## ART and AZD5582 treatment of BLT mice and NHPs

ART was administered to BLT mice using irradiated Teklad chow containing emtricitabine (1,500 mg kg<sup>-1</sup>), tenofovir disoproxil fumarate (1,560 mg kg<sup>-1</sup>) and raltegravir (600 mg kg<sup>-1</sup>) (Research Diets). Both the AZD5582 and the vehicle control (10% sterile captisol dissolved in sterile distilled water) were prepared fresh for each administration. AZD5582-2HCl was obtained from ChemieTek and dissolved at 5 mg ml<sup>-1</sup> in sterile distilled water (Gibco, Life Technologies) containing 10% captisol  $\beta$ -cyclodextrine sulfobutyl ethers sodium salts (Cydex Pharmaceuticals). AZD5582 was administered by intraperitoneal injection at a dose of 3 mg kg<sup>-1</sup>.

Rhesus macaques were treated with a potent 3-drug ART regimen initiated 56 days after infection that consisted of 2 reverse transcriptase inhibitors, tenofovir disoproxil fumarate (5.1 mg ml<sup>-1</sup>) and emtricitabine (40 mg ml<sup>-1</sup>) plus the integrase inhibitor dolutegravir (2.5 mg ml<sup>-1</sup>). ART was administered once daily at 1 ml kg<sup>-1</sup> body weight via the subcutaneous route. Peak plasma viral load (measured by the standard assay) and plasma viral load before LRA intervention (as measured by the ultrasensitive assay) were controlled for when allocating macaques into experimental groups. In total, 12 rhesus macaques were treated with AZD5582 and 9 rhesus macaques served as controls. AZD5582 was infused weekly intravenously at 0.1 mg kg<sup>-1</sup>. Three rhesus macaques received 3 doses of AZD5582 and were euthanized 48 h after the last dose. Nine rhesus macaques received 10 doses of AZD5582 and 3 macaques were euthanized 48 h after the last dose. Among the control rhesus macaques, 4 macaques received a weekly placebo infusion, 2 macaques were euthanized 48 h after 3 infusions and 2 macaques were euthanized 48 h after 10 infusions. The remaining five control macaques received ART only.

## Resting CD4<sup>+</sup> T cell enrichment

Human resting CD4<sup>+</sup> T cells were enriched from total cells isolated from BLT mouse tissues as follows. Each tissue from each mouse was processed individually and then tissues were pooled for immunomagnetic sorting. Each pooled tissue was first enriched for human cells with the EasySep Mouse/Human Chimera Kit (Stem Cell Technologies) and then for resting CD4<sup>+</sup> T cells with a human custom selection kit that included the following antibodies: CD8, CD14, CD16, CD19, CD20, CD36, CD56, CD123, glycophorin A, CD66b, CD25 and HLA-DR (all Stem Cell Technologies). Flow cytometry was performed before and after the enrichment to confirm the efficacy of the sort and purity of the sorted samples.

Resting CD4<sup>+</sup> T cells from rhesus macaques were isolated from the peripheral blood, bone marrow, lymph node and spleen. Before sorting, CD4<sup>+</sup> T cells were enriched using magnetic beads and column purification (Miltenyi Biotec). Enriched CD4<sup>+</sup> T cells were then stained with previously determined volumes of the following fluorescently conjugated antibodies: CD3 AF700 (clone SP34-2), CD8 APC-Cy7 (clone SK1), CD69 PE-CF594 (clone FN50), HLA-DR PerCP-Cy5.5 (clone G46-6) (all from BD Bioscience) and CD4 BrilliantViolet (BV)650 (clone OKT4) and CD25 PE-Cy7 (clone BC96) (both from Biolegend). Resting CD4<sup>+</sup> T cells were defined as CD3<sup>+</sup>CD4<sup>+</sup>CD8<sup>-</sup>CD69<sup>-</sup>CD25<sup>-</sup>HLA-DR<sup>-</sup>. Sorting was performed on a FACSAria LSR II (BD Biosciences) equipped with FACSDiva software.

## Plasma cytokine and chemokine analysis

Human plasma cytokine and chemokine analysis in BLT mice was performed by the University of North Carolina, Chapel Hill Center for AIDS Research Virology Core Laboratory. Plasma was tested undiluted in single wells using a Milliplex MAP kit (Millipore, HCYTMAG-60K-PX41) on a Luminex MAGPIX instrument. The following markers were tested: EGF, eotaxin, FGF-2, FLT-3L, fractalkine, G-CSF, GM-CSF, GRO, IFN $\alpha$ 2, IFN $\gamma$ ,

IL-1 $\alpha$ , IL-1 $\beta$ , IL-1RA, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-12p40, IL-12p70, IL-13, IL-15, IL-17A, IP-10, MCP-1, MCP-3, MDC, MIP-1 $\alpha$ , MIP-1 $\beta$ , PDGF-AA, PDGF-AB/BB, RANTES, sCD40L, TGF $\alpha$ , TNF, TNF $\beta$  and VEGF.

Rhesus macaque plasma levels of proinflammatory cytokines and chemokines were evaluated using the NHP MSD V-Plex assay systems developed by MesoScale Discovery. The two validated kits used were the V-PLEX Plus Proinflammatory Panel 1 NHP Kit (K15056D) that evaluates IFN $\gamma$ , IL-10, IL-1 $\beta$ , IL-2, IL-6 and IL-8, and a custom Chemokine NHP Kit that evaluates IP-10, MCP-1 and MIP-1 $\beta$  (K15055G). The manufacturer-provided protocol was followed with a few modifications. As the plasma samples were infectious, Triton X-100 at a final concentration of 1% was added not only to the subject samples but also to the assay diluents provided with the MSD kit before use. Calibration standards were reconstituted according to the protocol provided with the assay diluent that had the Triton X-100 added to it so that all standards and samples had the same components present. Plasma (60  $\mu$ l) was diluted twofold with assay diluent and then 50  $\mu$ l of the diluted sample was added to the assay plate along with 50  $\mu$ l of the calibrator according to the manufacturer's protocol. The plate was then covered and incubated on a shaker for 2 h. After the 2-h incubation, plates were washed at least 6 $\times$  with the supplied wash buffer and then 25  $\mu$ l of detection antibodies was added according to the kit protocol. Subsequently, the plate was covered and incubated on a shaker for 2 h at room temperature. After the detection incubation was finished, plates were then washed 6 $\times$  and 150  $\mu$ l of 2 $\times$  read buffer was added to each well and plates were analysed on the Sector s600 MSD plate reader. Data analysis was performed using the MSD Discovery Workbench analysis software.

## Cell-associated HIV RNA quantification in BLT mice

For tissue RNA analysis, RNA was extracted using QIAamp viral RNA columns (Qiagen) according to the manufacturer's protocol including an optional treatment with RNase-free DNase and analysed using one-step reverse-transcriptase qPCR (ABI custom TaqMan Assays-by-Design)<sup>46</sup>. Known quantities of HIV *gag* RNA standards were run in parallel, creating a standard curve for HIV *gag* and sample RNA was quantified by extrapolation from the standard curve. All samples were run and analysed on an ABI 7500 Fast Real-Time PCR System (Applied Biosystems). Owing to the relatively low number of human cells found in the brain, HIV RNA levels were quantified for only three vehicle control- and three AZD5582-treated mice.

## Immunohistochemical analysis in BLT mice

Tissues for immunohistochemical analysis were collected from BLT mice and fixed in 10% formalin for 16–24 h at 4  $^{\circ}$ C. Samples were then embedded in paraffin, cut into 5- $\mu$ m sections and mounted onto poly-L-lysine-coated glass slides. After paraffin removal, antigen retrieval (DIVA Decloaker, Biocare Medical) and blocking of nonspecific immunoglobulin-binding sites (Background Sniper, Biocare Medical), tissue sections were stained with anti-cIAP1 antibody (R&D Systems) overnight at 4  $^{\circ}$ C. To detect cIAP1, sections were probed with a goat-on-rodent HRP polymer (Biocare Medical) and developed with diaminobenzidine (ImmPact DAB Peroxidase Substrate, Vector Laboratories). As an isotype control, tissue sections were stained with polyclonal goat IgG (R&D Systems) negative control antibodies. Tissue sections were imaged with a Nikon Eclipse Ci microscope using Nikon Elements BR software (v.4.30.01) and a Nikon Digital Sight DS-Fi2 camera.

## Serum chemistry analysis in mice

After treatment with AZD5582 for 24 h, serum chemistry analysis in female 20-week-old BALB/cj mice (The Jackson Laboratory) was performed by the University of North Carolina, Chapel Hill Animal Histopathology Core Laboratory of the Lineberger Comprehensive Cancer Center. All clinical chemistry was performed on an Alfa Wassermann Vet Axcel analyser using Alfa Wassermann reagents. Automated assays for each analyte were as follows and performed according to the

manufacturer's instructions: alkaline phosphatase, Tietz-optimized Bowers and McComb assay; aspartate aminotransferase, Henry modification of Karmen's assay; alanine aminotransferase, Henry modification of Wroblewski and LaDue assay; creatinine, Jaffe reaction; blood urea nitrogen, enzymatic assay; albumin, Doumas and Briggs modification of the bromocresol green dye method; amylase, based on the use of chromogenic 2-chloro-*p*-nitrophenol linked with maltotriose; calcium, calcium-arsenazo assay phosphorus, based on the method of Daly and Ertingshausen with modifications by Armador and Urban; total bilirubin, Walters and Gerarde modification of the DMSO method; total protein, modification of Weichselbaum's biuret reagent.

### Immunophenotyping of BLT mice by flow cytometry

Immunophenotyping was performed on peripheral-blood samples collected longitudinally and at study end point, on blood and mononuclear cells isolated from the tissues of BLT mice. All flow cytometry data were collected on either BD LSR Fortessa or BD FACSCanto instruments using BD FACSDiva software (v.6.1.3) and data were analysed with FlowJo software (v.10.4.2). Antibodies used for longitudinal monitoring of human cells in peripheral blood include anti-CD45 APC (clone HI30, BD Biosciences), anti-CD3 FITC (clone HIT3a, BD Biosciences), anti-CD4 PE (clone RPA-T4, BD Biosciences) and anti-CD8 PerCP (clone SK1, BD Biosciences). Flow cytometry gating for the expression of lineage-specific antigens on human leukocytes was performed as follows. Step 1, forward- and side-scatter properties were used to set a live-cell gate. Step 2, live cells were then analysed for expression of the human pan-leukocyte marker CD45. Step 3, human leukocytes were then analysed for human CD3<sup>+</sup> T cells. Step 4, T cells were analysed for human CD4 and CD8 expression (Supplementary Fig. 2a).

At collection, peripheral blood and cells isolated from each individual tissue were stained with antibodies to detect human CD45, CD3, CD4, CD8, CD38 (anti-CD38 APC, clone HB7, BD Biosciences) and HLA-DR (anti-HLA-DR PE, clone TU36, BD Biosciences) to assess T cell activation. The following gating strategy was used. Step 1, forward- and side-scatter properties were used to set a live cell gate. Step 2, live cells were then analysed for expression of the human pan-leukocyte marker CD45. Step 3, human leukocytes were subsequently analysed for human CD3<sup>+</sup> T cells. Step 4, T cells were analysed for human CD4 and CD8 expression. Step 5, either CD4<sup>+</sup> or CD8<sup>+</sup> T cells were analysed for the expression of CD38 and HLA-DR (Supplementary Fig. 2b).

To determine the success of the sort and the purity of the sorted human resting CD4<sup>+</sup> T cells, the following antibodies were used to analyse pre-sort and post-sort samples: anti-CD3 BV421 (clone UCHT1, BD Biosciences), anti-HLA-DR PerCP (clone L243, BD Biosciences), anti-CD4 BV605 (clone RPA-T4, BD Biosciences), anti-CD8 APC-Cy7 (clone SK1, BD Biosciences), anti-CD25 APC (clone 2A3, BD Biosciences) and anti-CD45 V500 (clone HI30, BD Biosciences). Antibodies used as isotype controls: anti-mouse IgG1k APC (clone MOPC-21, BD Biosciences), anti-mouse IgG2 $\alpha$ k PerCP (clone X39, BD Biosciences), anti-mouse IgG1k PE (clone MOPC-21, BD Biosciences), anti-mouse IgG1k PE-Cy7 (clone MOPC-21, BD Biosciences) and anti-mouse IgG2 $\alpha$ k FITC (clone G155-178, BD Biosciences). Flow cytometry gating was performed as follows. Step 1, live cells were gated based on forward scatter and side scatter. Step 2, human haematopoietic cells were gated based on expression of human CD45. Step 3, human T cells were gated based on expression of CD3. Step 4, T cell subsets were gated based on the expression of human CD8 and CD4. Step 5, expression of activation markers (CD25 and HLA-DR) were analysed on the surface of CD4<sup>+</sup> T cells (Supplementary Fig. 2c).

### NHP sample collection and processing

EDTA-anticoagulated blood samples were collected regularly and used for a complete blood count, routine chemical analysis and immunostaining. Plasma was separated by centrifugation within 1 h of phlebotomy. At the end of the studies, tissue samples were collected, including lymph nodes (21 rhesus macaques), spleen (10 rhesus

macaques) and bone marrow (11 rhesus macaques). After 2 washes in RPMI and removal of connective and fat tissues, lymph nodes were ground using a 70- $\mu$ m cell strainer. PBMCs and bone-marrow mononuclear cells were prepared by density-gradient centrifugation. CD4<sup>+</sup> T cells were negatively selected from fresh or frozen cell suspensions using magnetically labelled microbeads and subsequent column purification according to the manufacturer's protocol (Miltenyi Biotec).

### AZD5582 pharmacokinetics in NHPs

Three healthy male rhesus macaques of Indian origin (aged 6–7 years, 10.5–13.0 kg) were used for the study. Fasted rhesus macaque received 0.1 mg kg<sup>-1</sup> of AZD5582 in 10% captisol and  $\leq$ 5% DMSO (dose volume 0.25 ml kg<sup>-1</sup> at 0.40 mg ml<sup>-1</sup>), filtered during administration (0.22- $\mu$ m PES in-line filter, Millex), via a 30-min saphenous vein infusion (infusion pump, Harvard Apparatus). Blood from femoral venipuncture or saphenous catheters was collected to obtain samples for plasma pharmacokinetics (K2 EDTA microtainers spun at 13,000 rpm for 5 min to obtain plasma), flow cytometry (Cytochex tubes (Streck)) and PBMC isolation (CPT blood collection tubes, BD Biosciences) processed according to the manufacturers' instructions and a dry pellet of cells was snap-frozen on dry ice and stored at -80 °C at various times after administration. AZD5582 was extracted from macaque plasma samples with an isotopically labelled internal standard (rilpivirine-d6) using protein precipitation. The compound was then eluted from a Waters Atlantis T3 (50 mm  $\times$  2.1 mm, 3  $\mu$ m particle size) analytical column under reverse-phase conditions and detected on an AB Sciex API-5000 triple quadrupole mass spectrometer under Turbolonspray mode. Standards were prepared in singlet and quality controls in duplicate, and a calibration curve was generated using a weighted 1/(x<sup>2</sup>) linear regression of the concentration (x) versus the analyte:internal standard peak area ratio (y). Concentrations of quality controls and study samples were calculated from this calibration curve using Sciex Analyst Software (v.1.6.2). The acceptance criterion of the assay was  $\pm$ 25% of the nominal concentration for standards and quality controls, and the quantifiable range was 2–2,000 ng ml<sup>-1</sup>.

### Ex vivo analysis of AZD5582 activity in NHP cells

Splenocytes from rhesus macaques were initially processed as above and frozen in FBS and 10% DMSO. Cryopreserved splenocytes were thawed and exposed to a range of concentrations of AZD5582 for either 48 h continuously or for 1 h after which the drug was removed by washing and cells were further cultured for 47 h. In preparation for western blot analysis, cells were lysed on ice with intermittent vortex mixing using NP40 Cell Lysis Buffer (Invitrogen, FNN0021) supplemented with appropriately diluted 10 $\times$  Protease Inhibitor Cocktail (Sigma, P-2714), 1 mM PMSF (Sigma, 93482) and 1 mM DTT (Sigma, 43816) to create a final concentration of 1 mM each in the complete lysis buffer. The lysed cells were then centrifuged at 13,000 rpm for 10 min at 4 °C and then supernatants (soluble fraction lysates) were removed and stored at -80 °C or used immediately for protein concentration and western blot analysis.

### Flow cytometry assay for p100 protein

Whole blood collected in Cyto-Chex blood collection tubes (Myriad RBM) was added to ACK lysis buffer to remove red-blood cells, washed with PBS and resuspended in a staining cocktail that included anti-CD3 BV421 (clone SP34-2, BD Biosciences), anti-CD16 BV605 (clone SG8, BD Biosciences), anti-CD4 BV711 (clone L200, BD Biosciences), anti-CD14 BV786 (clone M5E2, BD Biosciences), anti-CD123 PerCP-Cy5.5 (clone 7G3, BD Biosciences), anti-CD20 PE-CF594 (clone 2H7, BD Biosciences), anti-CD8 PE-Cy7 (clone SK1, BD Biosciences), anti-CD11c Alexa700 (clone 3.9, eBioscience) and anti-HLA-DR APC-Cy7 (clone L243, BD Biosciences). After surface-staining, cells were washed twice with PBS, permeabilized using the Cytotfix/Cytoperm kit (BD Biosciences, kit used as directed) and stained intracellularly with anti-p100 antibody

# Article

(clone EPR18756; Abcam). Cells were washed twice and stained with a secondary chicken anti-rabbit Alexa Fluor 488 antibody (Invitrogen). Samples were acquired using a four-laser Fortessa flow cytometer (BD Biosciences) and analysed with FlowJo software (version 9.7.6, TreeStar).

## SIV *env* sequencing and analysis

To generate *env* cDNA, reverse transcription of viral RNA was performed using SuperScript III reverse transcriptase according to the manufacturer's directions (Invitrogen) and the gene-specific primer SIVenvR1 5'-TGTAATAATCCCTTCCAGTCCCCC-3'. Single-genome amplification of SIV *env* was performed by serially diluting this cDNA in independent PCR reactions to identify a dilution in which amplification occurred in <30% of the total number of reactions. PCR amplification was performed with 1× PCR buffer, 2 mM MgSO<sub>4</sub>, 0.2 mM of each deoxynucleoside triphosphate, 0.2 μM of each primer and 0.025 U μl<sup>-1</sup> Platinum Taq High Fidelity polymerase (Invitrogen) in a 20-μl reaction. First-round PCR was performed with primer SIVenvF1 5'-CCTC-CCTCCAGGACTAGC-3' and antisense primer SIVenvR1 under the following conditions: 1 cycle of 94 °C for 2 min, 35 cycles at 94 °C for 15 s, 55 °C for 30 s and 68 °C for 5 min, followed by a final extension of 68 °C for 10 min. Next, 1 μl from the first-round PCR product was added to a second-round PCR reaction that included the sense primer SIVenvF2 5'-TATAATAGACATGGAGACACCCTTGAGGGAGC-3' and antisense primer SIVenvR2 5'-ATGAGACATRTCTATTGCCAATTGTA-3' performed under the same conditions used for first-round PCR, but with a total of 45 cycles. Correctly sized amplicons were identified by agarose gel electrophoresis and directly sequenced with second-round PCR primers and nine SIV-specific primers using BigDye Terminator technology (Applied Biosystems). To confirm PCR amplification from a single template, chromatograms were manually examined for multiple peaks, indicative of the presence of amplicons that resulted from PCR-generated recombination events, Taq polymerase errors or multiple variant templates. Alignment and phylogenetic trees were implemented in Geneious Prime 2019 (Biomatters) using the Muscle algorithm and the neighbour-joining method with the Tamura–Nei genetic distance model, respectively.

## Cell-associated SIV RNA and DNA quantification in NHPs

Cell-associated SIV RNA and DNA were measured simultaneously in resting CD4<sup>+</sup> T cells isolated from peripheral blood, spleen, bone marrow and lymph nodes (52,279–500,000 cells) or in total CD4<sup>+</sup> T cells isolated from peripheral blood or lymph nodes (100,000–500,000 cells), lysed in RLT+ Buffer (Qiagen), and stored at –80 °C. Nucleic acids were extracted using the AllPrep DNA/RNA mini kit (Qiagen) according to the manufacturer's recommendations with an on-column DNase digestion step. Cell-associated DNA quantification of SIV<sub>mac239</sub> *gag* DNA was performed on the extracted cell-associated DNA by qPCR using a 5' nuclease (TaqMan) assay with SIV *gag* primers and normalized to the rhesus macaque albumin gene, as previously described<sup>47</sup>. For cell-associated RNA quantification, RNA was reverse-transcribed using the High Capacity cDNA Reverse Transcription kit (ThermoScientific) and random hexamers. SIV *gag* and the rhesus macaque *CD4* gene were quantified by qPCR of the resultant cDNA using Taqman Universal Mastermix II (ThermoScientific). The CD4 primer and probe sequences were Rh-CD4-F 5'-ACATCGTGGTGC TAGCTTCCAGA-3', Rh-CD4-R 5'-AAGTGTAAGGCGAGTGGGAAGGA-3' and Rh-CD4-probe 5'-AGGC-CTCCAGCACAGTCTATAAGAAAGAGG-3'. The means of two replicate wells were used in all analyses. Samples with undetectable SIV DNA or RNA were assigned a level of 1 copy per million cells or million CD4 RNA copies, respectively, for display purposes.

## SIV quantitative viral outgrowth assay

Replication-competent SIV reservoirs were measured by the Viral Reservoir Core Laboratory of the Emory Center for AIDS Research. Latently infected cells were quantified using a limiting dilution culture assay in

which CD4<sup>+</sup> T cells—enriched from lymph node or spleen cells using magnetic beads and column purification (Miltenyi Biotec)—were co-cultured with CEMx174 cells in fivefold serial dilutions ranging from 5 × 10<sup>6</sup> cells per well to 4 × 10<sup>5</sup> cells per well. The cells were cultured in RPMI containing 10% FBS and 100 U ml<sup>-1</sup> IL-2 (Sigma). The ratio of target cells added was 4:1 for the two highest dilutions. A constant number of 1 × 10<sup>6</sup> CEMx174 cells was added to all other wells. The cultures were split every 7 days, and fresh medium was added. After 21 days, the growth of virus was detected by RT–qPCR. SIV RNA was isolated from 400 μl of culture supernatant using the Zymo viral RNA isolation kit (Zymo Research). DNase treatment was performed using a RQ1 RNase-free DNase kit (Promega). A one-step RT–qPCR targeting SIV *gag* was performed using an Applied Biosystems 7500 Real Time PCR System (Applied Biosystems) and the Taqman Fast Virus 1-step Master Mix (Thermo Scientific) for qRT–PCR with the following primers and probe: SIVgagFwd 5'-GCAGAGGAGGAAATTACCCAGTAC-3', SIVgagRev 5'-CAATTTTACCCAGGCATTTAATGTT-3' and SIVgag probe 5'-6FAM-TGTCCACCTGCCATTAAGCCCGA-3IBFQ-3'. The frequencies of infected cells were determined by the maximum-likelihood method<sup>48</sup> and were expressed as infectious units per million CD4<sup>+</sup> T cells.

## RNA-sequencing analysis of NHP cells

RNA-sequencing (RNA-seq) analysis was conducted at the Yerkes Non-human Primate Genomics Core Laboratory ([http://www.yerkes.emory.edu/nhp\\_genomics\\_core/](http://www.yerkes.emory.edu/nhp_genomics_core/)). RNA was purified from 50,000 peripheral blood- or lymph-node-derived CD4<sup>+</sup> T cells, which were purified by flow cytometry and lysed in 350 μl of RLT buffer at –80 °C, using Qiagen Micro RNEasy columns, and RNA quality was assessed using an Agilent Bioanalyzer. Then, 2 ng of total RNA was used as input for mRNA amplification using 5' template switch PCR with the Clontech SMART-seq v4 Ultra Low Input RNA kit according to the manufacturer's instructions. Amplified mRNA was fragmented and appended with dual-indexed bar codes using Illumina NexteraXT DNA library preparation kits. Libraries were validated by capillary electrophoresis on an Agilent 4200 TapeStation, pooled and sequenced on an Illumina HiSeq 3000 using 100-bp single reads at an average depth of 25 million reads.

## RNA-seq statistical analyses

RNA-seq data were mapped to the MacaM (v.7.8) assembly of the Indian rhesus macaque genome<sup>49</sup> (available at <https://www.unmc.edu/rhesusgenechip/index.htm>) and alignment was performed with STAR (v.2.5.2) using the annotation as a splice junction reference. Transcripts were annotated using the MacaM (v.7.8.2) annotation. Transcript abundance was estimated within STAR using the htseq-count algorithm and differential-expression analyses were performed using DESeq2. For significance testing using DESeq2, all of the samples from three- and ten-dose rhesus macaques were grouped together, separately for peripheral blood and lymph nodes, and compared with their pre-treatment samples. Genes determined to be differentially expressed by DESeq2 (adjusted  $P < 0.05$  and fold change  $> \pm 1.5$ ) were tested for enrichment of molecular pathways using the DAVID database<sup>50</sup> (<https://david.ncifcrf.gov/>) and by using GSEA with the desktop module (<https://www.broadinstitute.org/gsea/>). Gene sets for GSEA analysis were selected from the MSigDB database (<http://software.broadinstitute.org/gsea/msigdb/index.jsp>) and from the previously described 'blood transcriptome modules'<sup>51</sup>. Heat maps, DAVID enrichment bar charts and GSEA enrichment plots were generated with the R (v.3.5.0) package ggplot2. Principal component analysis was performed using Partek Genomics Suite software (v.6.6) using a covariance matrix.

## Immunophenotyping of NHPs by flow cytometry

Multicolour flow cytometry analysis was performed on whole blood and lymph-node mononuclear cells using predetermined optimal concentrations of the following fluorescently conjugated monoclonal antibodies: CD3 APC-Cy7 (clone SP34-2), Ki-67 AF700 (clone B56), HLA-DR

PerCP-Cy5.5 (clone G46-6), CCR5 APC (clone 3A9), CCR7 FITC (clone A20), CD45RA PE-Cy7 (clone 5H9) and CD62L PE (clone SK11) (all from BD Biosciences); CD8 BV711 (clone RPA-T8), CD4 BV650 (clone OKT4), CD95 BV605 (clone DX2) and PD-1 BV421 (clone EH12.2H7) (all from Biolegend) and CD28 PE-Cy5.5 (clone CD28.2) (from Beckman Coulter). Flow cytometry acquisition and analysis of samples was performed on at least 100,000 events on an LSRII flow cytometer driven by the FACSDiva software package (BD Biosciences). Analyses of the acquired data were performed using FlowJo software (Tree Star, v.10.0.4).

### ELISPOT for SIV-specific IFN $\gamma$ production in NHPs

IFN $\gamma$  production was evaluated after the stimulation of PBMCs with SIV<sub>mac239</sub> Gag and SIV<sub>mac239</sub> Env peptide pools. The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: SIV<sub>mac239</sub> Env Peptide Pool and the SIV<sub>mac239</sub> Gag Peptide Pool. To analyse IFN $\gamma$  expression, we used the Monkey Elispot kit from MABTECH. The manufacturer's instructions were followed as provided with the exception that concanavalin A (final concentration of 2.5  $\mu\text{g ml}^{-1}$ ) was used as the positive control agent. ELISPOT plates were blocked for 30 min with 100  $\mu\text{l}$  of RPMI 1640 supplemented with 10% FBS. PBMCs ( $4 \times 10^5$  per well) were incubated with 1  $\mu\text{g ml}^{-1}$  of DMSO, SIV<sub>mac239</sub> Gag Peptide Pool or SIV<sub>mac239</sub> Env Peptide Pool for 18 h before running the assay. For each plate, concanavalin A was used for the positive control wells. Samples were run in duplicate.

### Intracellular cytokine staining

PBMCs from five ART-suppressed SIV-infected rhesus macaques were thawed. Pre-treatment with AZD5582 was performed for 1 h at 100 nM. After two washes, phorbol 12-myristate 13-acetate (PMA) and ionomycin were added for 1 h at 500 ng  $\text{ml}^{-1}$  and 10  $\mu\text{g ml}^{-1}$ , respectively. DMSO treatment controls were prepared in parallel. After 1 h, brefeldin-A and Golgi stop solution were added following the manufacturer's recommendations (BD GolgiStop Protein Transport Inhibitor, BD Biosciences). Cells were incubated at 37 °C, 5% CO<sub>2</sub> in R10 for 6–8 h before staining with the following antibodies: CD3 APC-Cy7 (clone SP34-2), CD8 BV711 (clone RPA-T8), CD4 BV650 (clone OKT4) (all from Biolegend), and IFN $\gamma$  PE (clone B7), TNF AF700 (clone Mab11) and IL-2 BV605 (clone MQ1-17H12) (all from BD Biosciences). Flow cytometry acquisition and analysis of samples was performed on at least 100,000 events using an LSRII flow cytometer driven by the FACSDiva software package (BD Biosciences). Analyses of the acquired data were performed using FlowJo (TreeStar, v.10.0.4) and simplified presentation of incredibly complex evaluations (SPICE, v.6.0)<sup>52</sup> software.

### Proliferation assay

PBMCs from five ART-suppressed SIV-infected rhesus macaques were thawed and labelled with CellTrace Violet Proliferation Kit according to the manufacturer instructions (Molecular Probes). Cells were plated in R10 in 96-well round-bottom plates at 4 million cells per ml. Pre-treatment with AZD5582 was performed for 1 h at 100 nM. After two washes, PMA and ionomycin were added for 1 h at 500 ng  $\text{ml}^{-1}$  and 10  $\mu\text{g ml}^{-1}$ , respectively. DMSO treatment controls were prepared in parallel. Cells were incubated at 37 °C, 5% CO<sub>2</sub> in R10 supplemented with 20 IU  $\text{ml}^{-1}$  IL-2. After 5 days, cells were stained with CD3 APC-Cy7 (clone SP34-2), CD8 BV711 (clone RPA-T8), CD4 BV650 (clone OKT4) (all from Biolegend) and analysed by flow cytometry on an LSRII instrument. Data were analysed using FlowJo software (TreeStar, v.10.0.4).

### Statistical analysis

Statistical analyses were performed using GraphPad Prism Software (v.6 or v.7).  $P \leq 0.05$  was considered statistically significant. At least three samples were used for each group, the minimum to achieve statistical significance. No statistical methods were used to predetermine sample size. Investigators were not blinded to group allocations or when assessing outcomes. In some instances, cells were pooled from individual

humanized mice for each tissue and experimental group for the isolation of resting CD4<sup>+</sup> T cells (Fig. 2d and Extended Data Fig. 2). To test the statistical significance of the differences that we observed in PCR data in BLT mouse brains in Fig. 2e and grouped tissue PCR data in Fig. 2d, unpaired two-sided Student's *t*-tests were used. To test the statistical significance of the differences that we observed in PCR data in PBMCs and female reproductive tract in Fig. 2e, T cell activation markers in Supplementary Table 5 and plasma cytokines and chemokines in Supplementary Table 6, we used unpaired two-sided Mann–Whitney *U*-tests. To assess the statistical significance of the differences observed over time in serum enzymes in Supplementary Table 4, we used paired nonparametric Wilcoxon matched-pairs signed-rank tests. To test the statistical significance of the differences in activation marker levels expressed on T cells from rhesus macaques in Fig. 4d, e, as well as the ultrasensitive plasma viral load results in Extended Data Fig. 8b and the ex vivo proliferation assay in Extended Data Fig. 10g, we used Wilcoxon matched-pairs signed-rank tests. To assess the statistical significance of the differences observed between SIV DNA and SIV RNA levels in resting or total CD4<sup>+</sup> T cells from rhesus macaques in Fig. 3e and Extended Data Fig. 6b, respectively, as well as for the quantitative viral outgrowth results in Extended Data Fig. 6c, we used unpaired two-sided Mann–Whitney *U*-tests. Unpaired two-sided Mann–Whitney *U*-tests were also used to compare rhesus macaques with stable compared with increased viraemia in Extended Data Fig. 8c.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

Source Data for Figs. 1–4, Extended Data Figs. 1, 2, 4, 6–10 and Supplementary Tables 4–6 are provided with the paper. Gene-expression data are available at the Gene Expression Omnibus (GEO) repository (accession number GSE141546 and GSE142774). Any other data are available from corresponding authors on reasonable request.

- Archin, N. M. et al. Expression of latent HIV induced by the potent HDAC inhibitor suberoylanilide hydroxamic acid. *AIDS Res. Hum. Retroviruses* **25**, 207–212 (2009).
- Keedy, K. S. et al. A limited group of class I histone deacetylases acts to repress human immunodeficiency virus type 1 expression. *J. Virol.* **83**, 4749–4756 (2009).
- Trumble, I. M. et al. SLDAssay: a software package and web tool for analyzing limiting dilution assays. *J. Immunol. Methods* **450**, 10–16 (2017).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
- Denton, P. W. et al. Antiretroviral pre-exposure prophylaxis prevents vaginal transmission of HIV-1 in humanized BLT mice. *PLoS Med.* **5**, e16 (2008).
- Denton, P. W. et al. One percent tenofovir applied topically to humanized BLT mice and used according to the CAPRISA 004 experimental design demonstrates partial protection from vaginal HIV infection, validating the BLT model for evaluation of new microbicide candidates. *J. Virol.* **85**, 7582–7593 (2011).
- Reed, L. J. & Muench, H. A simple method of estimating fifty per cent endpoints. *Am. J. Epidemiol.* **27**, 493–497 (1938).
- Palesch, D. et al. Short-term pegylated interferon  $\alpha 2a$  treatment does not significantly reduce the viral reservoir of simian immunodeficiency virus-infected, antiretroviral therapy-treated rhesus macaques. *J. Virol.* **92**, e00279-18 (2018).
- Hansen, S. G. et al. Addendum: immune clearance of highly pathogenic SIV infection. *Nature* **547**, 123–124 (2017).
- Li, H. et al. Envelope residue 375 substitutions in simian–human immunodeficiency viruses enhance CD4 binding and replication in rhesus macaques. *Proc. Natl Acad. Sci. USA* **113**, E3413–E3422 (2016).
- Krisko, J. F., Martinez-Torres, F., Foster, J. L. & Garcia, J. V. HIV restriction by APOBEC3 in humanized mice. *PLoS Pathog.* **9**, e1003242 (2013).
- Cartwright, E. K. et al. CD8<sup>+</sup> lymphocytes are required for maintaining viral suppression in SIV-infected macaques treated with short-term antiretroviral therapy. *Immunity* **45**, 656–668 (2016).



48. Rosenbloom, D. I. S., Hill, A. L., Laskey, S. B. & Siliciano, R. F. Re-evaluating evolution in the HIV reservoir. *Nature* **551**, E6–E9 (2017).
49. Zimin, A. V. et al. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol. Direct* **9**, 20 (2014).
50. Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44–57 (2009).
51. Li, J. et al. Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. *Blood* **124**, 3636–3645 (2014).
52. Roederer, M., Nozzi, J. L. & Nason, M. C. SPICE: Exploration and analysis of post cytometric complex multivariate datasets. *Cytometry* **79A**, 167–174 (2014).

**Acknowledgements** We thank Garcia and Chahroudi laboratory members, the Animal Histopathology & Laboratory Medicine Core at the University of North Carolina-Chapel Hill (UNC-CH), which is supported in part by an NCI Center Core Support Grant (5P30CA016086-41) to the UNC Lineberger Comprehensive Cancer Center and technicians from the Department of Comparative Medicine at UNC-CH; the HIV/STD Laboratory Core and the Clinical Pharmacology and Analytical Chemistry Core of the UNC Center for AIDS Research (CFAR) (P30 AI050410); D. Hazuda, B. Howell and S. Barrett (Merck & Co.) for assistance with ART-containing chow; Yerkes Animal and Research Resources; the Children's Healthcare of Atlanta and Emory University Pediatric Flow Cytometry Core, Emory CFAR Translational Virology and Reservoir Cores, the Quantitative Molecular Diagnostics Core of the AIDS and Cancer Virus Program, Frederick National Laboratory, as well as GSK for tenofovir disoproxil fumarate, emtricitabine and dolutegravir. This work was supported by the National Institutes of Allergy and Infectious Diseases (NIAID)(AI123010, AI096113, AI11899, AI117851 and P30 AI050410), and Mental Health (NIMH) (MH108179). This work was supported by the Emory Consortium for Innovative AIDS Research in Nonhuman Primates (UM1 AI124436), amfAR (109353-59-RGRL), the Yerkes National Primate Research Center (P51 OD011132) and the Translational Virology and Reservoir Cores of the Center for AIDS Research at Emory University (P30 AI050409). Research was also supported by Qura Therapeutics and by CARE, a Martin Delaney Collaboratory (1UM1AI126619-01) of the NIAID, NINDS, NIDA and NIMH. By the Natural Science Foundation of Guangdong Province, China (2016A030310108), UNC-South China STD Research Training Center (1D43TW009532) and Chinese National Key Technologies R&D

Program for the 13th Five-year Plan (2017ZX10202101003). Federal funds were used for this research from the National Cancer Institute, NIH (contracts HHSN261200800001E and 75N91019D0002). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

**Author contributions** J.V.G., A.C., G.S., R.M.D., A.W. and D.M.M. conceived and designed the studies. G.C.S., D.M.I., S.R.W., S.D.F., C.G., E.P.B. and R.G.F. assessed in vitro AZD5582 activity. M.K., Z.W., J.H.B., D.F. and C.D.J. performed the in vitro transcriptomic study. D.F. and J.-P.R. contributed clinical specimens. C.C.N., P.T.H., C.D., N.J.S. and B.L. performed mouse experiments. R.A.S. and C.G.C. performed the qPCR analysis of HIV RNA levels in mouse samples. R.A.C. performed the immunohistochemical analysis. N.M.A. isolated human resting CD4<sup>+</sup> T cells. C.C.N. analysed mouse data. M.M., A.D.B. and S.J. performed the monkey experiments. G.C.S., J.D., S.M. and R.M.D. conducted pharmacokinetic studies in NHP. C.M. and N.S. isolated resting CD4<sup>+</sup> T cells from macaques. T.H.V. supervised the qPCR analyses of SIV RNA and DNA levels in monkey samples. M.M. and A.D.B. performed flow cytometry and analysed monkey data. G.K.T., A.A.U. and H.W. performed the macaque RNA-seq analyses and S.E.B. analysed these data. C.M.F. and B.F.K. performed *env* sequencing analyses. J.D.L. supervised the ultrasensitive plasma viral loads. G.C.S. and C.G. analysed AZD5582 target engagement in human, mouse and macaque samples and J.D. and R.M.D. analysed the data. C.G. performed the ELISPOTs. C.C.N., M.M., A.W., A.C. and J.V.G. wrote the manuscript with input from all authors.

**Competing interests** The authors declare no competing interests.

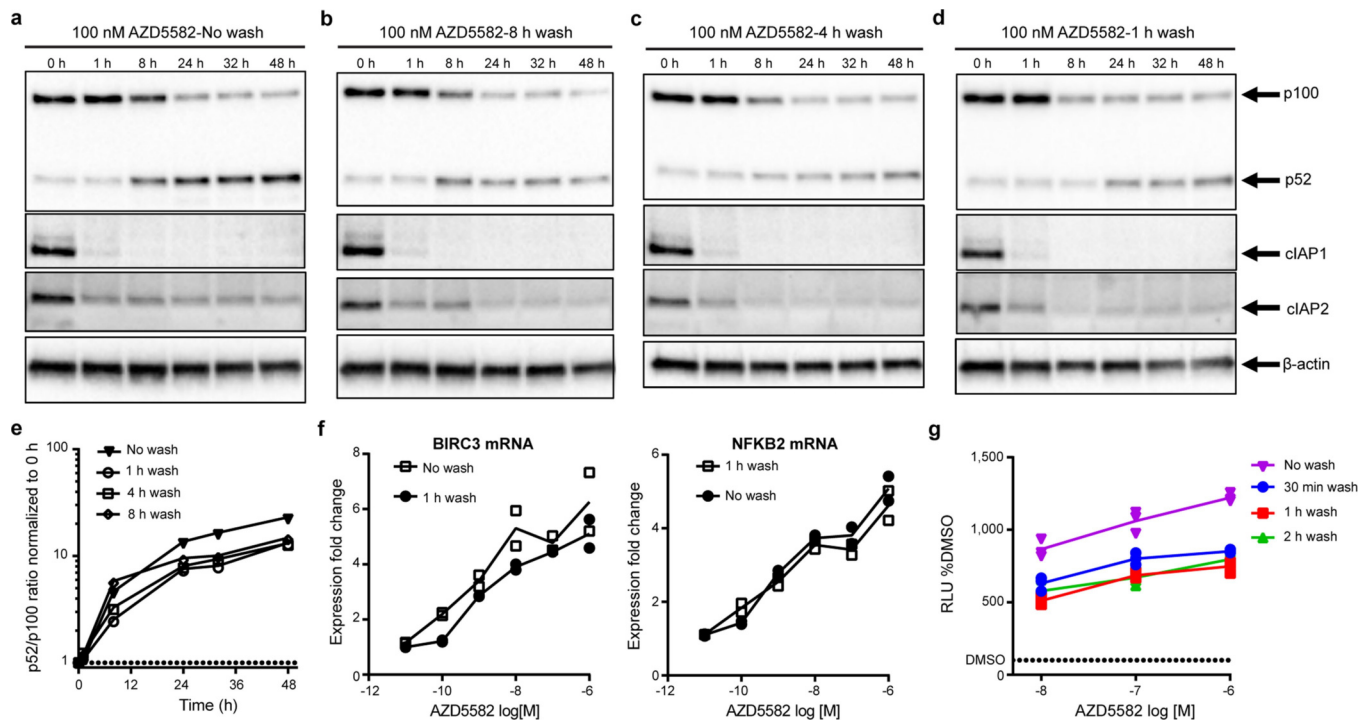
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1951-3>.

**Correspondence and requests for materials** should be addressed to R.M.D., A.C. or J.V.G.

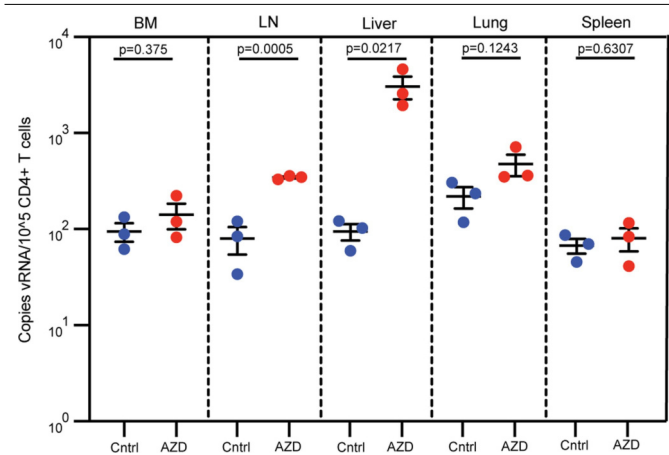
**Peer review information** *Nature* thanks Mathias Lichterfeld and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

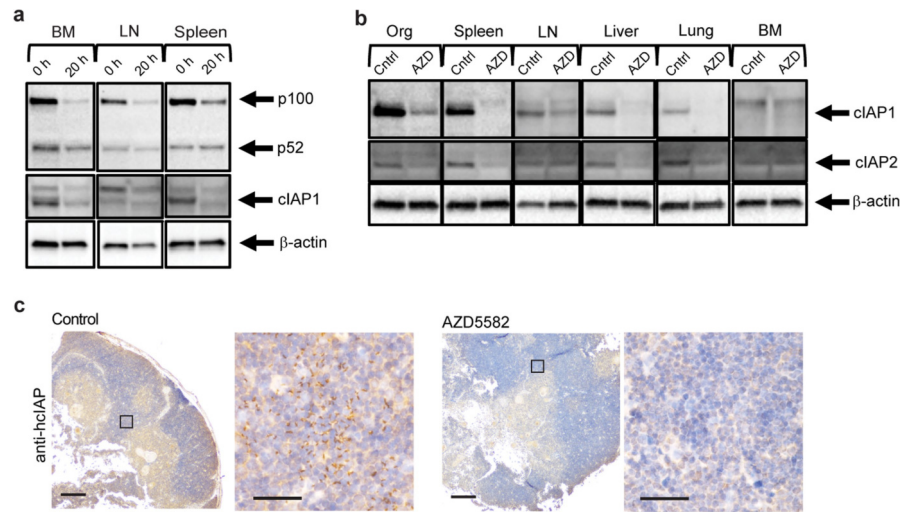


**Extended Data Fig. 1 | Short-duration exposure to AZD5582 activates the ncNF-κB pathway. a–d**, Isolated total human CD4<sup>+</sup> T cells treated with 100 nM AZD5582 and then either not washed (**a**) or washed three times with PBS 8 h (**b**), 4 h (**c**) or 1 h (**d**) after treatment. Whole-cell lysates were then analysed by immunoblot for components of the ncNF-κB pathway. **e**, Densitometry analysis of the ratio of p52 to p100 from the pulse-wash assay immunoblots. Points represent values for the densitometric ratio from one western blot, representative of several independent experiments. **a–e**, Entire set representative of 1 experiment with 5 replicates of the 1-h wash condition conducted. **f**, Fold induction of ncNF-κB target gene expression in isolated

CD4<sup>+</sup> T cells from an uninfected donor treated with the indicated concentrations of AZD5582 and either washed after 1 h or not washed, and subsequently cultured for 24 h as measured by RT-qPCR. Points represent two technical replicates and lines represent the mean. The data presented are representative of three independent experiments. **g**, DMSO-normalized induction of luciferase activity from the Jurkat reporter model after exposure to AZD5582 (10, 100 or 1,000 nM) for 30 min (blue), 1 h (red), 2 h (green) or continued exposure (purple). Points represent three replicates in one assay run, representative of two independent experiments. Lines represent the mean of the three replicates. For gel source data, see Supplementary Fig. 1.

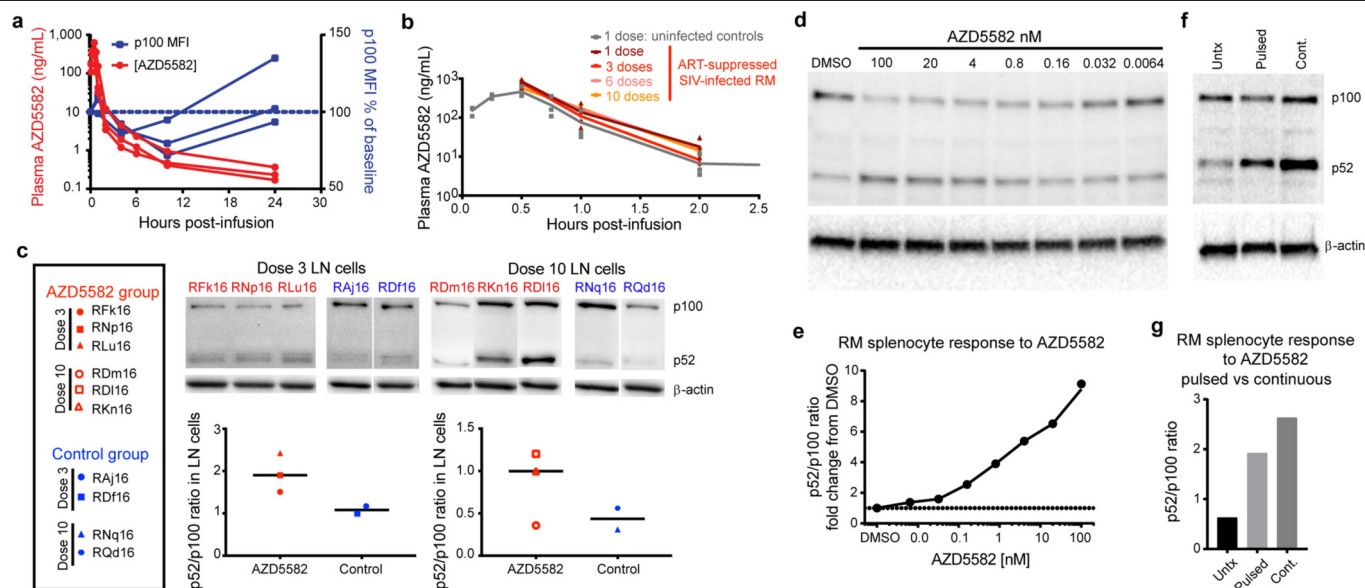


**Extended Data Fig. 2 | AZD5582 induces HIV RNA expression in resting CD4<sup>+</sup> T cells from tissues of HIV-infected, ART-suppressed BLT mice.** HIV RNA levels in resting CD4<sup>+</sup> T cells isolated from the bone marrow, lymph nodes, liver, lung and spleen of control (Cntrl, blue circles) or AZD5582-treated (AZD, red circles) mice (cells pooled from  $n=6$  mice per group for each tissue) were analysed in triplicate. Statistical significance was determined with a two-sided Student's  $t$ -test. The mean fold increase in viral RNA levels in resting CD4<sup>+</sup> T cells from tissues in the experiment shown in Fig. 2 was 12.1 ( $\pm 3.7$ ), whereas in this figure it was 8.4 ( $\pm 6.0$ ). These values were not statistically different ( $P=0.4286$ , two-sided Mann-Whitney  $U$ -test). Data are mean  $\pm$  s.e.m.



**Extended Data Fig. 3 | AZD5582 ex vivo target engagement. a,** Western blot analysis of p100, p52 and cIAP1 protein levels in cells isolated from the bone marrow, lymph node and spleen of BLT mice before and 20 h after ex vivo treatment with AZD5582. Loading control, β-actin. Representative of two experiments. **b,** cIAP expression in resting CD4<sup>+</sup> T cells isolated from the thymus, spleen, lymph nodes, liver, lung and bone marrow. Loading control, β-actin. Representative of two experiments. **c,** cIAP expression in the thymic

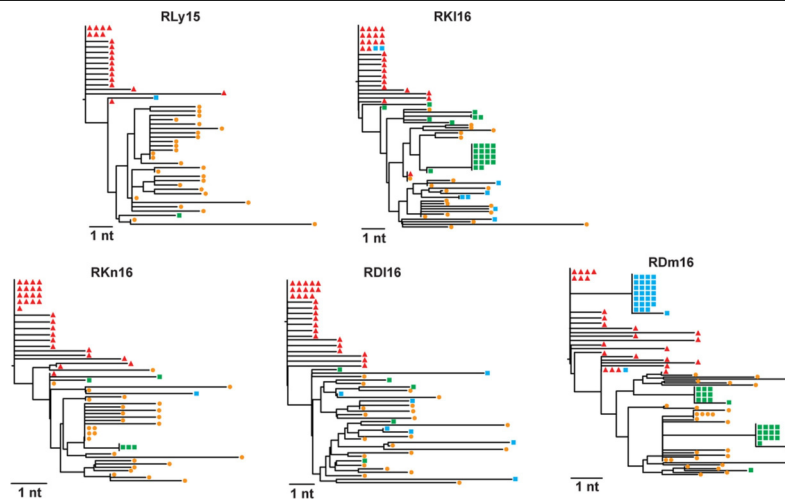
organoid of HIV-infected ART-suppressed BLT mice 48 h after the administration of vehicle control or AZD5582 (three control and one AZD5582-treated analysed). Positive cells, brown. Imaged at 4× and 40× magnifications; scale bars, 100 μm (4×) and 50 μm (40×). Boxes in the 4× images indicate regions corresponding to images at 40× magnification. For gel source data, see Supplementary Fig. 1.



**Extended Data Fig. 4 | Pharmacokinetic and pharmacodynamic assessment of AZD5582 in rhesus macaques.** **a**, AZD5582 ( $0.1 \text{ mg kg}^{-1}$ ) was administered to healthy rhesus macaques ( $n=3$ ) by intravenous infusion. Plasma concentrations of AZD5582 (left y-axis) are shown for the indicated time points. Flow cytometry was used to measure intracellular p100 levels, shown as the geometric mean fluorescence intensity (gMFI) in  $\text{CD4}^+$  T cells and plotted as the percentage of baseline of p100 gMFI (right y-axis). **b**, Plasma concentrations of AZD5582 after one (dark red), three (red), six (pink) or ten (orange) doses in six SIV-infected ART-treated rhesus macaques (RM) and after one dose in three uninfected control rhesus macaques (grey). Individual values are shown as symbols. **c**, Western blot analyses of inactive p100 and active p52 forms of NF- $\kappa$ B2 in lymph-node mononuclear cells collected 48 h after the third or tenth dose of AZD5582 in SIV-infected ART-suppressed rhesus macaques (red;  $n=3$  for both the 3-dose and 10-dose groups) or at equivalent time points for

placebo controls (blue;  $n=2$  for both the 3-dose and 10-dose groups). Immunoblots are shown in the top panels and densitometry analyses of the p52:p100 ratios are shown in the bottom panels. The line represents the median. **d**, Cryopreserved control rhesus macaque splenocytes were treated with the indicated concentrations of AZD5582 for 48 h, then p100/p52 levels were analysed by western blotting to measure engagement of the NF- $\kappa$ B pathway. **e**, DMSO-normalized densitometric p52:p100 ratio versus the AZD5582 concentration. For **d**, **e**, the experiments were performed in duplicate. **f**, Cryopreserved rhesus macaque splenocytes were exposed to DMSO alone (Untx), 100 nM AZD5582 washed off after 1 h and cultured for 47 h (Pulsed) or continuous 100 nM AZD5582 for 48 h (Cont.), after which the cells were studied by western blot for p100 and p52 levels. **g**, Densitometric p52:p100 ratio. For **f**, **g**, data represent a single experiment. For gel source data, see Supplementary Fig. 1.

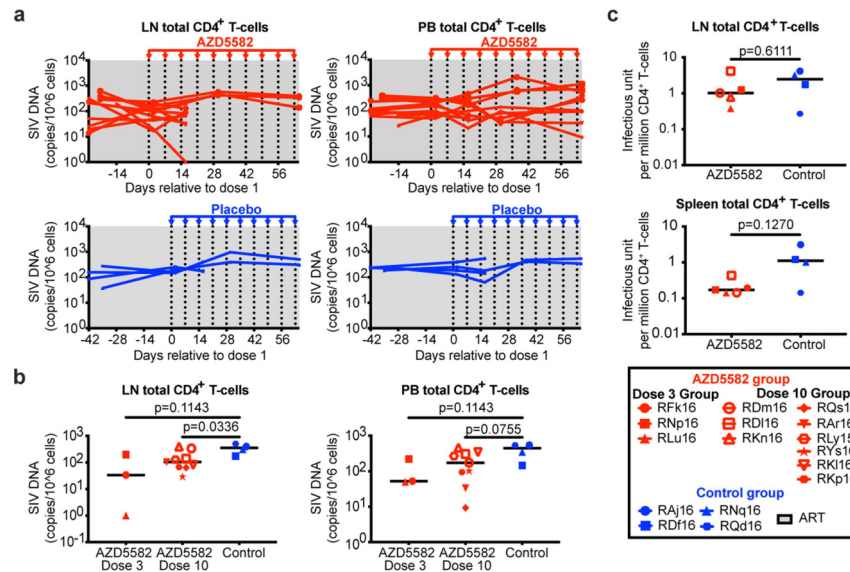




**Extended Data Fig. 5 | Phylogenetic trees based on full *env* sequencing.**

Plasma virus was sequenced from four time points per macaque ( $n = 5$ ): near peak viraemia (2 weeks after infection; red), immediately before ART (8 weeks after infection; orange) and two time points of on-ART viraemia during

AZD5582 treatment (green and blue). All sequences per macaque were phylogenetically analysed and the resulting phylogenetic trees are shown for each macaque. The horizontal bar indicates the genetic distance. nt, nucleotide.



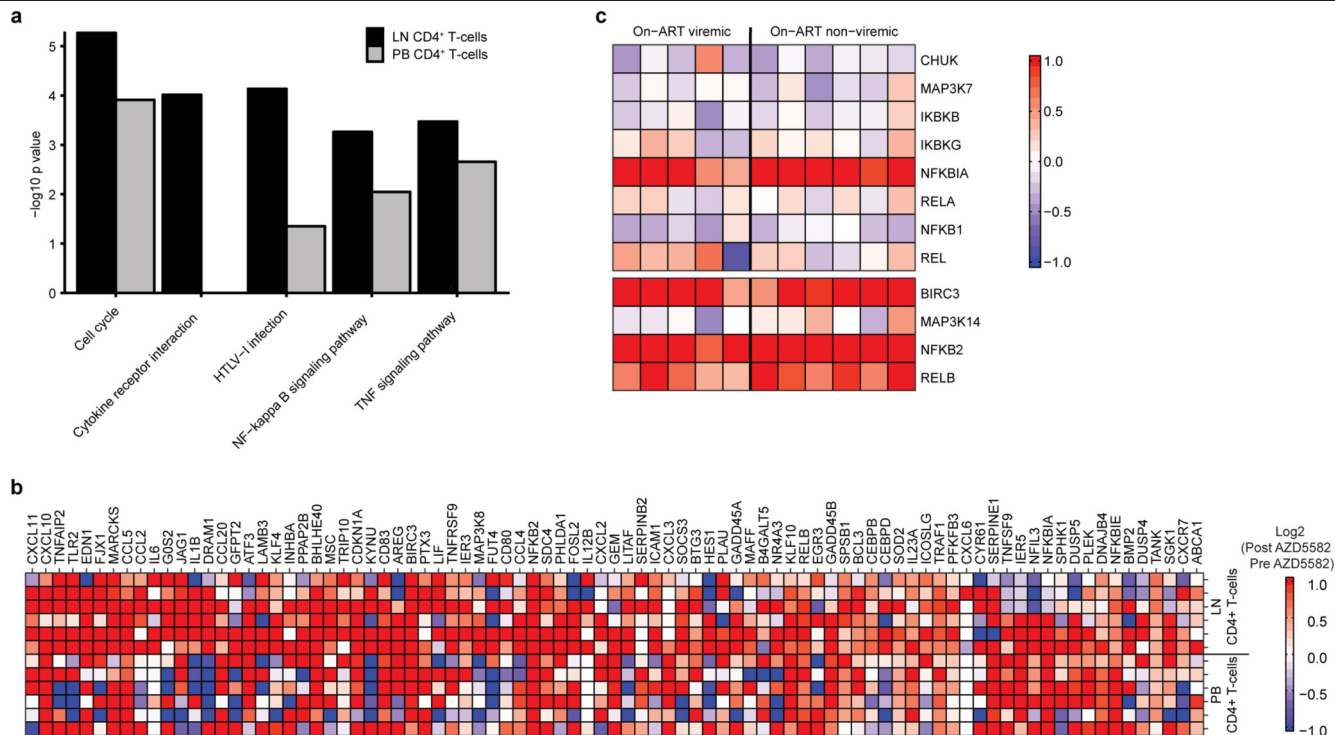
**Extended Data Fig. 6 | SIV DNA levels in total CD4<sup>+</sup> T cells and replication-competent reservoir size in ART-suppressed SIV-infected rhesus macaques.**

**a**, Longitudinal assessment of cell-associated SIV DNA levels in total CD4<sup>+</sup> T cells isolated from peripheral blood and lymph nodes of AZD5582-treated ( $n=12$ , red) and control ( $n=4$ , blue) ART-suppressed SIV-infected rhesus macaques. Grey shading represents the period of ART administration.

**b**, Comparison of cell-associated SIV DNA levels in total CD4<sup>+</sup> T cells isolated from the lymph nodes and peripheral blood of AZD5582-treated and control ART-suppressed SIV-infected rhesus macaques. Total CD4<sup>+</sup> T cells were analysed from AZD5582-treated rhesus macaques 48 h after receiving 3 doses (lymph nodes and peripheral blood,  $n=3$ ) or 10 doses (lymph nodes and peripheral blood,  $n=9$ ) of AZD5582. Total CD4<sup>+</sup> T cells were analysed from

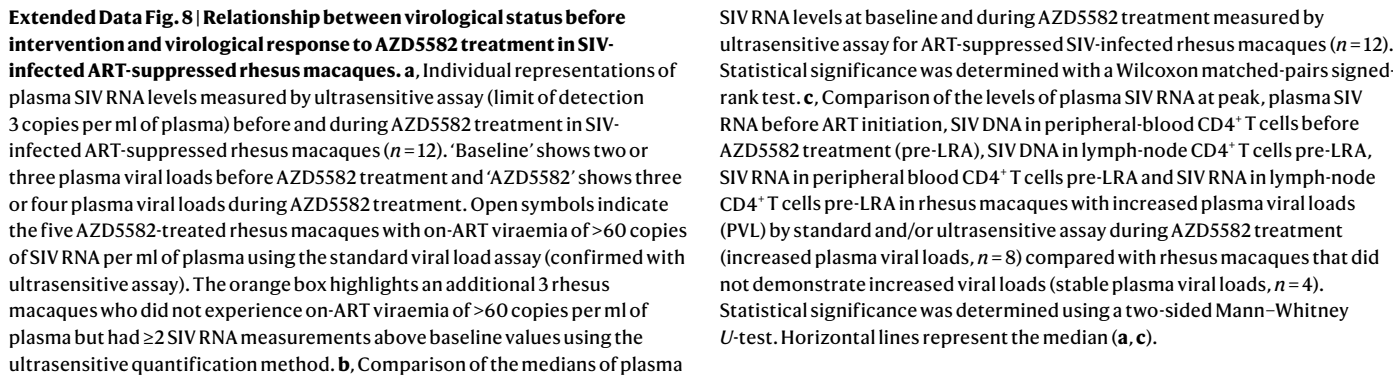
placebo-control rhesus macaques (lymph nodes and peripheral blood,  $n=4$ ) at equivalent time points. Open symbols indicate AZD5582-treated rhesus macaques with on-ART viraemia above 60 copies per ml of plasma. Statistical significance was determined using a two-sided Mann-Whitney  $U$ -test.

**c**, Quantitative viral outgrowth assays were performed for AZD5582-treated rhesus macaques 48 h after receiving 3 doses (lymph nodes,  $n=2$ ; spleen,  $n=3$ ) or 10 doses (lymph nodes,  $n=3$ ; spleen,  $n=2$ ) of AZD5582. Quantitative viral outgrowth assays were performed from control rhesus macaques (lymph nodes and spleen,  $n=4$ ) at equivalent time points. Open symbols indicate AZD5582-treated rhesus macaques with on-ART viraemia. Statistical significance was determined with a two-sided Mann-Whitney  $U$ -test. Horizontal lines represent the median (**b**, **c**).

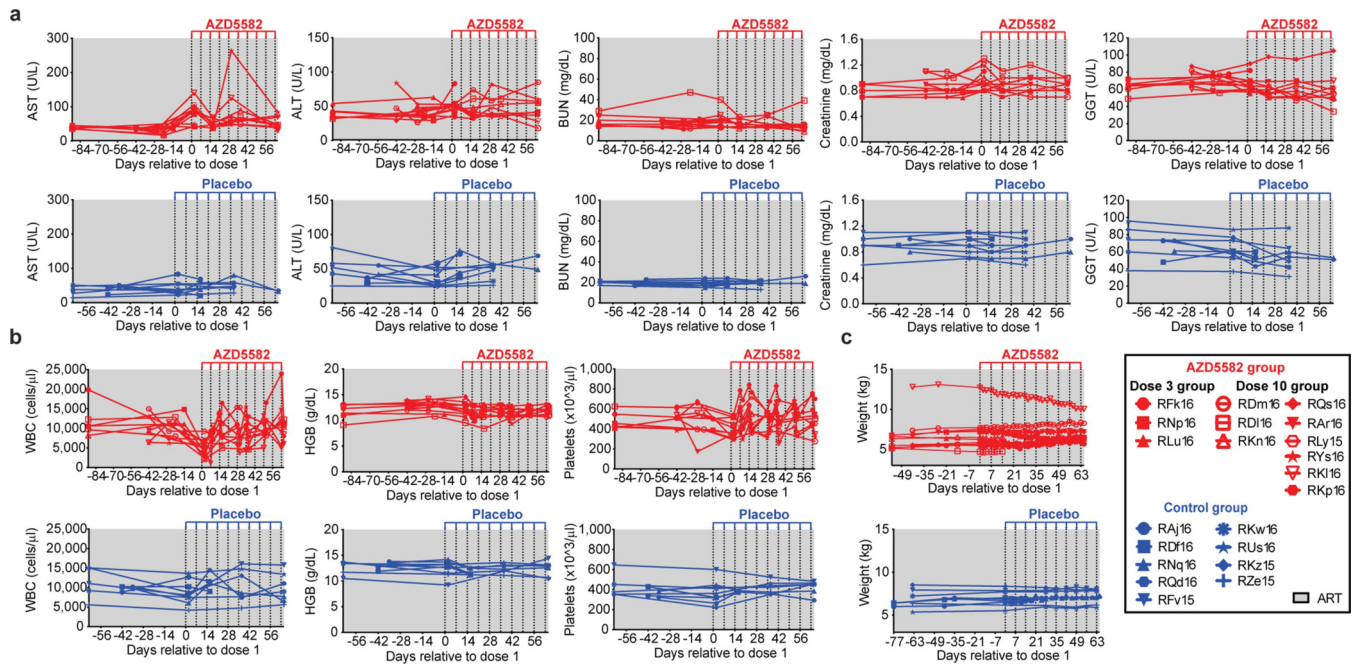


**Extended Data Fig. 7 | AZD5582-induced gene pathways and genes in SIV-infected ART-suppressed rhesus macaques.** **a**, DAVID analysis showing pathways significantly enriched in genes differentially expressed after AZD5582 treatment relative to baseline in CD4<sup>+</sup> T cells isolated from lymph nodes (black bars) and peripheral blood (grey bars).  $n=6$  for both lymph nodes and peripheral blood; for each,  $n=3$  for 3 doses of AZD5582 and  $n=3$  for 10 doses of AZD5582. **b**, Leading edge genes from the 'hallmark TNF signalling via NF- $\kappa$ B' pathway from MSigDB. Genes were identified in the leading edge of CD4<sup>+</sup> T cell samples from the lymph nodes before and after treatment with AZD5582 (shown in Fig. 4b). The contrast depicted is the fold change of each gene for each rhesus macaque's post-treatment sample relative to the pre-

treatment values for CD4<sup>+</sup> T cells from the lymph nodes (top) and the peripheral blood (bottom).  $n=6$  for both lymph nodes and peripheral blood; for each,  $n=3$  for 3 doses of AZD5582 and  $n=3$  for 10 doses of AZD5582. **c**, Heat map of cNF- $\kappa$ B (top) and ncNF- $\kappa$ B (bottom) pathway gene expression in rhesus macaques with (left,  $n=5$ ) or without (right,  $n=6$ ) on-ART viraemia of >60 copies per ml plasma. One rhesus macaque without on-ART viraemia was excluded from this analysis because of technical issues (higher than expected unmapped and multi-mapped reads, and lower than expected unique identified reads compared to the means). Colour scale,  $\log_2$ -transformed fold changes of post-treatment compared with pre-treatment values.



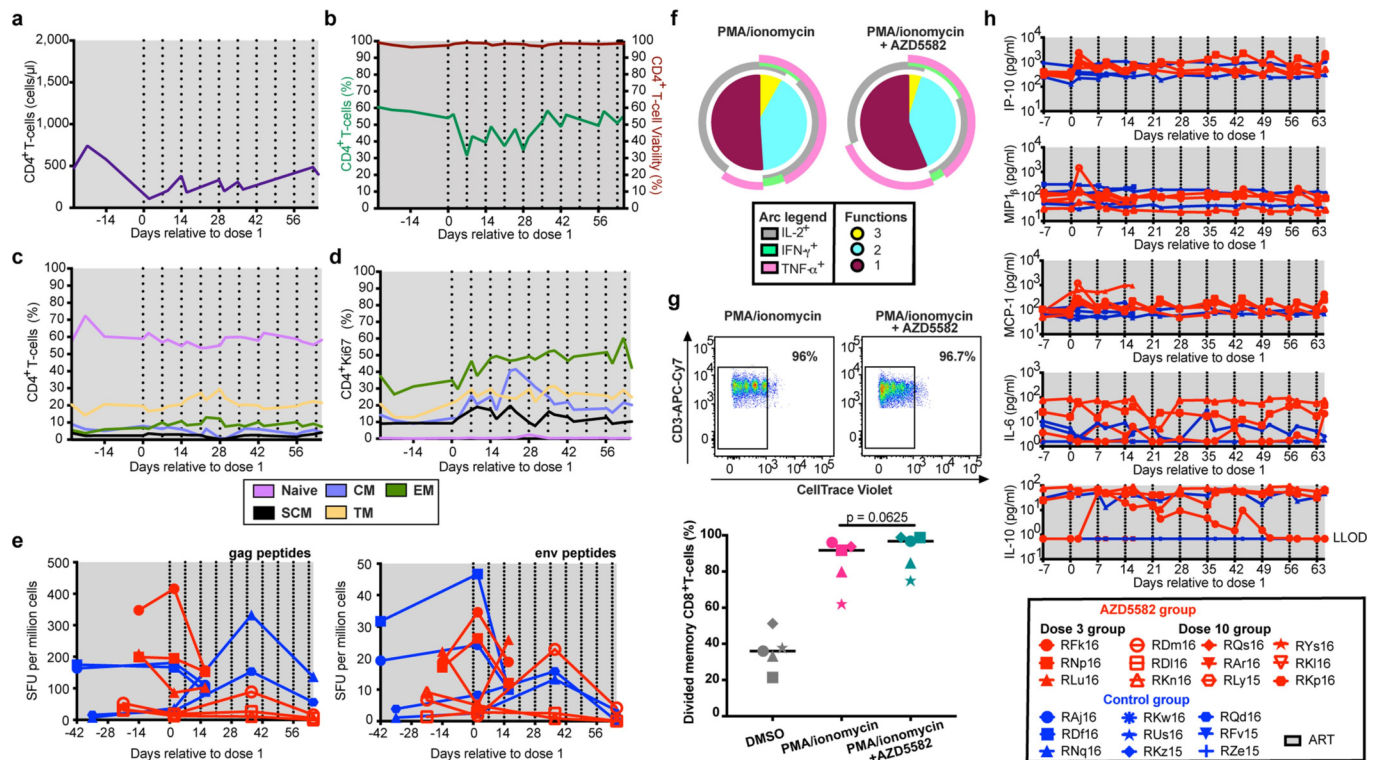
SIV RNA levels at baseline and during AZD5582 treatment measured by ultrasensitive assay for ART-suppressed SIV-infected rhesus macaques ( $n = 12$ ). Statistical significance was determined with a Wilcoxon matched-pairs signed-rank test. **c**, Comparison of the levels of plasma SIV RNA at peak, plasma SIV RNA before ART initiation, SIV DNA in peripheral-blood CD4<sup>+</sup> T cells before AZD5582 treatment (pre-LRA), SIV DNA in lymph-node CD4<sup>+</sup> T cells pre-LRA, SIV RNA in peripheral blood CD4<sup>+</sup> T cells pre-LRA and SIV RNA in lymph-node CD4<sup>+</sup> T cells pre-LRA in rhesus macaques with increased plasma viral loads (PVL) by standard and/or ultrasensitive assay during AZD5582 treatment (increased plasma viral loads,  $n = 8$ ) compared with rhesus macaques that did not demonstrate increased viral loads (stable plasma viral loads,  $n = 4$ ). Statistical significance was determined using a two-sided Mann–Whitney *U*-test. Horizontal lines represent the median (**a**, **c**).



**Extended Data Fig. 9 | AZD5582 can be safely administered in SIV-infected ART-suppressed rhesus macaques. a–c.** Longitudinal assessment of serum chemistries (a), complete blood counts (b) and weight (c) of SIV-infected ART-treated rhesus macaques treated with AZD5582 (red,  $n = 12$ ) compared with

controls (blue,  $n = 9$ ). Grey shading represents the period of ART administration. AST, aspartate aminotransferase; ALT, alanine aminotransferase; BUN, blood urea nitrogen; GGT,  $\gamma$ -glutamyltransferase; HGB, haemoglobin; WBC, white blood cell.





**Extended Data Fig. 10 | Immunological effect of AZD5582 on SIV-infected ART-suppressed rhesus macaques.** **a, b**, Longitudinal assessment of the count (**a**) and the frequency and viability (**b**) of CD4<sup>+</sup> T cells in SIV-infected ART-treated rhesus macaques during the period of AZD5582 treatment ( $n=12$ ). Lines represent median values. **c, d**, CD4<sup>+</sup> T cell naive and memory subset frequencies (**c**) and their expression of Ki-67 (**d**) in SIV-infected ART-treated rhesus macaques during the period of AZD5582 treatment ( $n=12$ ). Lines represent median values. **e**, SIV gag-specific (left) and SIV env-specific (right) CD8<sup>+</sup> IFN $\gamma$ <sup>+</sup> T cell responses in SIV-infected, ART-suppressed, AZD5582-treated (red,  $n=6$ ) and control (blue,  $n=4$ ) rhesus macaques. SFU, spot-forming units. **f**, Pie charts depicting the ability of memory CD8<sup>+</sup> T cells isolated from SIV-infected ART-suppressed control rhesus macaques ( $n=5$ ) to produce IFN $\gamma$ , IL-2 and/or TNF in response to stimulation with PMA and ionomycin in the absence

(left) or presence (right) of AZD5582 pre-treatment. **g**, Memory CD8<sup>+</sup> T cell proliferative response to stimulation with PMA and ionomycin with or without AZD5582 pre-treatment. Top, representative flow cytometry dot plots gated on memory CD8<sup>+</sup> T cells. Bottom, comparison between divided cells 5 days after stimulation in each group ( $n=5$  for each). Statistical significance was determined with a Wilcoxon matched-pairs signed-rank test. Horizontal lines represent the median. **h**, Longitudinal assessment of plasma levels of IP-10, MIP-1 $\beta$ , MCP-1, IL-6 and IL-10 by multiplex assay in SIV-infected ART-suppressed rhesus macaques treated with AZD5582 (red,  $n=6$ ) or control rhesus macaques (blue,  $n=4$ ). An additional four analytes (IFN $\gamma$ , IL-8, IL-1 $\beta$  and IL-2) were undetectable in all macaques. **a–e, h**, Grey shading represents the period of ART administration. Dashed lines represent AZD5582 or placebo infusions. LLOD, lower limit of detection.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection	Nikon Elements BR software (version 4.30.01), BD FACSDiva software (version 6.1.3), ChemiDoc MP Image Lab software (version 6.0.1, BioRad), QuantStudio3 Real-Time PCR System software (version 1.4.3, ThermoFisher)
Data analysis	FlowJo software (version 10), ChemiDoc MP Image Lab software (version 6.0.1, BioRad), QuantStudio3 (ThermoFisher), FASTQC (version 0.11.1), STAR (version 2.5.2), Salmon (version 0.7.2), GSEA (version 2.2.3), Go Analysis (Go Panther 11.1), Sciex Analyst software (version 1.6.2), Geneious software (Biomatters), Discovery Workbench software (version 4.0, Meso Scale Delivery), R (version 3.5.0), GraphPad Prism (versions 6 and 8), Applied Biosystems 7500 Real-Time PCR software (version 2.0.6), Simplified Presentation of Incredibly Complex Evaluations (SPICE) software (version 6.0), Partek Genomics Suite software (version 6.6)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability

The data generated are available from corresponding authors on reasonable request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For BLT mouse studies, no statistical methods were used to predetermine sample size. At least 3 animals were used for each experimental group, the minimum to achieve statistical significance. Based on our previous data on SIV-infected ART-treated rhesus macaques, with a sample size of at least 7, we would be able to detect a significant difference between pre- and post- AZD5582 treatment samples in the levels of plasma RNA at the 0.05 significance level with a power of 0.90.
Data exclusions	Data exclusion was applied only to one RNAseq analysis represented on the heatmap of extended data Figure 7. One RM without on-ART viremia was excluded from this analysis for technical issues (higher than expected unmapped and multi-mapped reads, and lower than expected unique identified reads compared to the means).
Replication	In Figure 1b, symbols represent technical replicates of DMSO-normalized reporter signal induced by a dose titration of a panel of mono- and bivalent SMACm in a Jurkat luciferase reporter model of HIV-1 latency with 48 h exposure. In Extended Data Fig. 1f, fold induction of ncNF-kB target gene expression was measured by quantitative RT-PCR. Points represent two technical replicates. The data presented are representative of three independent experiments. In Extended Data Fig. 1g, for DMSO-normalized induction of luciferase activity from the Jurkat reporter model after exposure to AZD5582, points represent three replicates in one assay run, representative of several independent experiments. All attempts at replication were successful. In two independent experiments, plasma viremia was observed following AZD5582 administration to HIV-infected, ART-suppressed BLT mice. The use of non-human primates precludes our ability to replicate experiments. Sample sizes were chosen to maximize the likelihood of detecting statistical differences.
Randomization	For the study that examined the impact of AZD5582 administration on plasma and tissue viremia in BLT mice during ART suppression, mice were randomized for assignment to either experimental or control groups using randomization software available at random.org. For RM studies, peak plasma viral load (measured by standard assay) and plasma viral load before LRA intervention (as measured by ultrasensitive assay) were controlled for when allocated animals into experimental groups.
Blinding	Investigators were not blinded to group allocations or when assessing outcomes. In some instances, cells were pooled from individual humanized mice for each tissue and experimental group for the isolation of resting CD4+ T cells (Fig. 2).

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	rat anti-mouse CD24-biotin (clone M1/69), BD Biosciences (Cat. # 557436), 10ug biotin-labeled rat anti-mouse CD24 antibody was adsorbed to 1mg streptavidin-labeled magnetic Dynabeads , <a href="https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/mouse/purified-rat-anti-mouse-cd24-m169/p/557436">https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/mouse/purified-rat-anti-mouse-cd24-m169/p/557436</a> anti-clAP1-unconjugated (clone EPR4673), Abcam (Cat. # 108361), 1:1,000 dilution, <a href="https://www.abcam.com/ciap1-antibody-epr4673-ab108361.html">https://www.abcam.com/ciap1-antibody-epr4673-ab108361.html</a> anti-p100/p52-unconjugated (clone 18D10), Cell Signaling Technology (Cat. # 3017), 1:1,000 dilution, <a href="https://www.cellsignal.com/products/primary-antibodies/nf-kb2-p100-p52-18d10-rabbit-mab/3017">https://www.cellsignal.com/products/primary-antibodies/nf-kb2-p100-p52-18d10-rabbit-mab/3017</a> anti-IkBα-unconjugated (clone 44D4), Cell Signaling Technology (Cat. # 4812), 1:1,000 dilution, <a href="https://www.cellsignal.com/products/primary-antibodies/ikba-44d4-rabbit-mab/4812">https://www.cellsignal.com/products/primary-antibodies/ikba-44d4-rabbit-mab/4812</a>
-----------------	--

anti-clAP2-unconjugated (clone E40), Abcam (Cat. # ab32059), 1:1,000 dilution, <https://www.abcam.com/ciap2-antibody-e40-ab32059.html>

anti-beta-actin-HRP (clone AC-15 ), Abcam (Cat. # ab49900), 1:30,000 dilution, <https://www.abcam.com/beta-actin-antibody-ac-15-hrp-ab49900.html>

anti-clAP1-unconjugated (goat polyclonal IgG), R&D Systems (Cat. # AF8181, Lot # KH50516111), 10 ug/ml dilution, [https://www.rndsystems.com/products/human-ciap-1-hiap-2-antibody\\_af8181](https://www.rndsystems.com/products/human-ciap-1-hiap-2-antibody_af8181)

anti-CD3-BV421 (clone SP34-2), BD Biosciences (Cat. # 562877), 1:250 dilution, <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-human-antibodies/cell-surface-antigens/bv421-mouse-anti-human-cd3-sp34-2/p/562877>

anti-CD16-BV605 (clone 3G8), BD Biosciences (Cat. # 563172), 1:50 dilution, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/bv605-mouse-anti-human-cd16-3g8/p/563172>

anti-CD4-BV711 (clone L200), BD Biosciences (Cat. # 563913), 1:50 dilution, <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-non-human-primate-antibodies/cell-surface-antigens/bv711-mouse-anti-human-cd4-l200/p/563913>

anti-CD14-BV786 (clone M5E2), BD Biosciences (Cat. # 563698), 1:50 dilution, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/hematopoietic-stem-cell-markers/human/negative-markers/bv786-mouse-anti-human-cd14-m5e2/p/563698>

anti-CD123-PerCP-Cy5.5 (clone 7G3), BD Biosciences (Cat. #558714), 1:25 dilution, <https://www.bdbiosciences.com/us/applications/research/b-cell-research/surface-markers/human/percp-cy55-mouse-anti-human-cd123-7g3/p/558714>

anti-CD20-PE-CF594 (clone 2H7), BD Biosciences (Cat. # 562295), 1:500 dilution, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/hematopoietic-stem-cell-markers/human/negative-markers/pe-cf594-mouse-anti-human-cd20-2h7/p/562295>

anti-CD8-PE-Cy7 (clone SK1), BD Biosciences (Cat. # 335787), 1:500 dilution, <https://www.bdbiosciences.com/us/reagents/research/clinical-research---ruo-gmp/single-color-antibodies/pe-cytrade7-mouse-anti-human-cd8-sk1/p/335787>

anti-CD11c-Alexa700 (clone 3.9), Ebioscience (Cat. # 50-112-9413), 1:50 dilution, <https://www.thermofisher.com/antibody/product/CD11c-Antibody-clone-3-9-Monoclonal/56-0116-42>

anti-HLA-DR-APC-Cy7 (clone L243), BD Biosciences (Cat. # 335796), 1:50 dilution, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/mesenchymal-stem-cell-markers-bone-marrow/human/negative-markers/apc-cytrade7-mouse-anti-human-hla-dr-l243/p/335796>

anti-p100-unconjugated (clone EPR18756), Abcam (Cat. # ab191594), 1:25 dilution, <https://www.abcam.com/nfkb-p100nfkb2-antibody-epr18756-ab191594.html>

anti-CD45-APC (clone HI30), BD Biosciences (Cat. # 555485), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/apc-mouse-anti-human-cd45-hi30/p/555485>

anti-CD3-FITC (clone HIT3a), BD Biosciences (Cat. # 555339), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/fits-mouse-anti-human-cd3-hit3a/p/555339>

anti-CD19-PE (clone HIB19), BD Biosciences (Cat. # 555413), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/clinical-research/immunology-reagents/blood-cell-disorders/surface-markers/human/pe-mouse-anti-human-cd19-hib19/p/555413>

anti-CD4-PerCP (clone SK3), BD Biosciences (Cat. # 347324), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/percp-mouse-anti-human-cd4-sk3-also-known-as-leu3a/p/347324>

anti-CD4-PE (clone RPA-T4), BD Biosciences (Cat. # 555347), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/pe-mouse-anti-human-cd4-rpa-t4/p/555347>

anti-CD8-PerCP (clone SK1), BD Biosciences (Cat. # 347314), 3 ul/test, <https://www.bdbiosciences.com/us/reagents/research/clinical-research---ruo-gmp/single-color-antibodies/percp-mouse-anti-human-cd8-sk1/p/347314>

anti-CD45-APC-Cy7 (clone 2D1), BD Biosciences (Cat. # 557833), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/apc-cy7-mouse-anti-human-cd45-2d1/p/557833>

anti-CD3-PE-Cy7 (clone SK7), BD Biosciences (Cat. # 557851), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/pe-cy7-mouse-anti-human-cd3-sk7-also-known-as-leu-4/p/557851>

anti-CD8-FITC (clone SK1), BD Biosciences (Cat. # 340692), 3 ul/test, <https://www.bdbiosciences.com/us/applications/clinical/blood-cell-disorders/asr-reagents/cd8-fits-sk1/p/340692>

anti-CD38-APC (clone HB7), BD Biosciences (Cat. # 340439), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/regulatory-t-cells/surface-markers/human/apc-mouse-anti-human-cd38-hb7/p/340439>

anti-HLA-DR-PE (clone TU36), BD Biosciences (Cat. # 555561), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/mesenchymal-stem-cell-markers-bone-marrow/human/negative-markers/pe-mouse-anti-human-hla-dr-tu36-also-known-as-t36-t36/p/555561>

anti-CD3-BV421 (clone UCHT1), BD Biosciences (Cat. # 562426), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/bv421-mouse-anti-human-cd3-ucht1-also-known-as-ucht-1/p/562426>

anti-HLA-DR-PerCP (clone L243), BD Biosciences (Cat. # 347364), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/mesenchymal-stem-cell-markers-bone-marrow/human/negative-markers/percp-mouse-anti-human-hla-dr-l243/p/347364>

anti-CD4-BV605 (clone RPA-T4), BD Biosciences (Cat. # 562658), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/bv605-mouse-anti-human-cd4-rpa-t4/p/562658>

anti-CD8-APC-Cy7 (clone SK1), BD Biosciences (Cat. # 557834), 3 ul/test, <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-human-antibodies/cell-surface-antigens/apc-cy7-mouse-anti-human-cd8-sk1/p/557834>

anti-CD25-APC (clone 2A3), BD Biosciences (Cat. # 340938), 3 ul/test, <https://www.bdbiosciences.com/us/applications/clinical/blood-cell-disorders/asr-reagents/cd25-apc-2a3/p/340938>

anti-CD45-V500 (clone H130), BD Biosciences (Cat. # 560777), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/v500-mouse-anti-human-cd45-hi30/p/560777>

anti-mouse IgG1k-APC (clone MOPC-21), BD Biosciences (Cat. # 555751), 3 ul/test, <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-human-antibodies/cell-surface-antigens/apc-mouse-igg1-isotype-control-mopc-21/p/555751>

anti-mouse IgG2ak-PerCP (clone X39), BD Biosciences (Cat. # 340765), 3 ul/test, <https://www.bdbiosciences.com/us/>

applications/clinical/blood-cell-disorders/asr-reagents/mouse-iggsub2asub-percp-x39/p/340765  
 anti-mouse IgG1k-PE (clone MOPC-21), BD Biosciences (Cat. # 559320), 3 ul/test, <https://www.bdbiosciences.com/us/applications/research/intracellular-flow/intracellular-antibodies-and-isotype-controls/anti-human-antibodies/pe-mouse-igg1-isotype-control-mopc-21/p/559320>  
 anti-mouse IgG1k-PE-Cy7 (clone MOPC-21), BD Biosciences (Cat. # 557872), 3 ul/test, <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-human-antibodies/cell-surface-antigens/pe-cy7-mouse-igg1-isotype-control-mopc-21/p/557872>  
 anti-mouse IgG2ak-FITC (clone G155-178), BD Biosciences (Cat. # 553456), 3 ul/test, <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-mouse-antibodies/cell-surface-antigens/fitc-mouse-igg2a-isotype-control-g155-178/p/553456>  
 anti-CD3-APC-Cy7 (clone SP34-2), BD Biosciences (Cat. # 557757), <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-non-human-primate-antibodies/cell-surface-antigens/apc-cy7-mouse-anti-human-cd3-sp34-2/p/557757>  
 anti-Ki-67-AF700 (clone B56), BD Biosciences (Cat. # 561277), <https://www.bdbiosciences.com/us/applications/research/intracellular-flow/intracellular-antibodies-and-isotype-controls/anti-human-antibodies/alexa-fluor-700-mouse-anti-ki-67-b56/p/561277>  
 anti-HLA-DR-PerCP-Cy5.5 (clone G46-6), BD Biosciences (Cat. # 560652), <https://www.bdbiosciences.com/us/applications/research/stem-cell-research/mesenchymal-stem-cell-markers-bone-marrow/human/negative-markers/percp-cy55-mouse-anti-human-hla-dr-g46-6/p/560652>  
 anti-CCR5-APC (clone 3A9), BD Biosciences (Cat. # 550856), <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/th-1-cells/surface-markers/human/apc-mouse-anti-human-cd195-3a9/p/550856>  
 anti-CD8-BV711 (clone RPA-T8), BD Biosciences (Cat. # 563677), <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-human-antibodies/cell-surface-antigens/bv711-mouse-anti-human-cd8-rpa-t8/p/563677>  
 anti-CD4-BV650 (clone OKT4), Biolegend (Cat. # 317436), <https://www.biolegend.com/en-us/products/brilliant-violet-650-anti-human-cd4-antibody-7786>  
 anti-PD-1-BV421 (clone EH12.2H7), Biolegend (Cat. # 329920), <https://www.biolegend.com/en-us/products/brilliant-violet-421-anti-human-cd279-pd-1-antibody-7191>  
 anti-CD3-AF700 (clone SP34-2), BD Bioscience (Cat. # 557917), <https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/anti-non-human-primate-antibodies/cell-surface-antigens/alexa-fluor-700-mouse-anti-human-cd3-sp34-2/p/557917>  
 anti-CD69-PE-CF594 (clone FN50), BD Bioscience (Cat. # 562617), <https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/regulatory-t-cells/surface-markers/human/pe-cf594-mouse-anti-human-cd69-fn50-also-known-as-fn-50/p/562617>  
 anti-CD25-PE-Cy7 (clone BC96), Biolegend (Cat. # 302612), <https://www.biolegend.com/en-us/products/pe-cy7-anti-human-cd25-antibody-1909>  
 anti-CD45RA-PE-Cy7 (clone5H9), BD Biosciences (Cat. # 561216), <https://www.bdbiosciences.com/us/applications/research/b-cell-research/surface-markers/non-human-primates/pe-cy7-mouse-anti-human-cd45ra-5h9/p/561216>  
 anti-CD62L-PE (clone SKI1), BD Biosciences (Cat. # 654666), <https://www.bdbiosciences.com/us/applications/clinical/blood-cell-disorders/asr-reagents/cd62l-pe-sk11-also-known-as-anti-leu-8/p/654666>  
 anti-CD95-BV605 (clone DX2), Biolegend (Cat. # 305627), <https://www.biolegend.com/en-us/products/brilliant-violet-605-anti-human-cd95-fas-antibody-8778>  
 anti-CD28-PE-Cy5.5 (clone CD28.2), Beckman Coulter (Cat. # B24027), <https://www.beckman.com/reagents/coulter-flow-cytometry/antibodies-and-kits/single-color-antibodies/cd28/b24027>

#### Validation

The specificity of the antibodies purchased from commercial sources (BD Biosciences, Abcam, Cell Signaling Technology, R&D Systems, Ebioscience, Biolegend, and Beckman Coulter) were validated by the manufacturer as noted on their website (links provided above for each antibody).

## Eukaryotic cell lines

### Policy information about cell lines

Cell line source(s)	Jurkat Clone E6-1 cells (American Type Culture Collection TIB-152), TZM-bl cells (NIH AIDS Reagent Repository) and HEK 293T cells (European Collection of Authenticated Cell Cultures)
Authentication	Cell lines were authenticated by morphological identification and virus susceptibility profiles.
Mycoplasma contamination	Cell lines were tested negative for mycoplasma by the supplier
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No commonly misidentified cell lines were used

## Animals and other organisms

### Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	BLT mice were constructed using 12-15 week old female NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice (NSG; The Jackson Laboratory, Bar Harbor, ME) mice. Female 20 week old BALB/cJ (The Jackson Laboratory, Bar Harbor, ME) were used for the serum chemistry analysis. Three healthy male rhesus macaques ( <i>Macaca mulatta</i> ) of Indian origin, age 6-7 years, were utilized for the AZD5582 pharmacokinetic study. Twenty-one male and female Mamu-B*08 and -B*17 negative rhesus macaques, age 3-6 years, were infected with SIVmac239 and treated with ART (Supplementary Table 7).
--------------------	---



Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	Mice were maintained under specific pathogen-free conditions by the Division of Comparative Medicine at the University of North Carolina, Chapel Hill. Mouse experiments were conducted in accordance with NIH guidelines for the housing and care of laboratory animals and in accordance with protocols reviewed and approved by the Institutional Animal Care and Use Committee at the University of North Carolina, Chapel Hill. Healthy Rhesus macaques for pharmacokinetic studies were housed at GlaxoSmithKline and all procedures were conducted in accordance with the GlaxoSmithKline Policy on the Care, Welfare, and Treatment of Laboratory Animals and were reviewed by the IACUC at GlaxoSmithKline. Rhesus macaques infected with SIV were housed at the Yerkes National Primate Research Center (Atlanta, GA) and treated in accordance with Emory University and Yerkes National Primate Research Center Institutional Animal Care and Use Committee regulations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Cells used in the QVOA and RNAseq experiments were obtained from participants stably suppressed on ART. At the time of sample donation, participants had a mean age of 43 years [range, 26-61 years], a mean duration on ART of 7 years [range, >6 months -22 years] and a mean CD4 count of 634 [range, 372-1364 cells/ $\mu$ l]. All participants were male, and 88% were Caucasians and 12% African American. Twenty-five percent of the participants were treated during acute infection and 75% during chronic infection.
Recruitment	Cells used in the QVOA assays were selected randomly across participants enrolled in a longitudinal reservoir measurement study and thus should not be subjected to self-selection bias. For the RNA seq experiments, cells from participants with a demonstrated increase in cell associated HIV RNA following ex-vivo exposure to AZD5582 were selected, potentially introducing a self-selection bias. However, given that these were global human gene expression measurements, we do not believe our results were impacted by this bias.
Ethics oversight	All human subjects samples were obtained under a specimen procurement protocol reviewed and approved by the University of North Carolina Biomedical Institutional Review Board and the McGill University Health Centre Ethical Review Board. Informed consent was obtained from all participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation	Sample preparation for the flow cytometric analysis of peripheral blood and tissues from humanized mice and rhesus macaques is detailed in the Methods section.
Instrument	Flow cytometry data was collected on BD LSRII, BD LSRFortessa, BD FACSAria LSR II, or BD FACSCanto instruments using BD FACSDiva software.
Software	Flow cytometry data was analyzed with FlowJo software.
Cell population abundance	Resting CD4+ T cells represented 0.12-9.48% (mean: 3.42%) of the total cell population. Prior to sorting of macaque resting CD4+ T cells by FACS, CD4+ T cells were enriched by magnetic bead selection. Post-sort purity was 97.8%.
Gating strategy	For the analysis of the frequency and phenotype of different human immune cell populations in the peripheral blood and tissues of BLT mice, an antibody specific for human CD45, a pan leukocyte marker, was used first to gate human leukocytes. Gates to define positive and negative populations were defined by isotype controls when appropriate.

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# PIK3CA variants selectively initiate brain hyperactivity during gliomagenesis

<https://doi.org/10.1038/s41586-020-1952-2>

Received: 18 December 2018

Accepted: 12 December 2019

Published online: 29 January 2020

Kwanha Yu<sup>1,10</sup>, Chia-Ching John Lin<sup>1,10</sup>, Asante Hatcher<sup>2</sup>, Brittney Lozzi<sup>1</sup>, Kathleen Kong<sup>1</sup>, Emmet Huang-Hobbs<sup>1</sup>, Yi-Ting Cheng<sup>1</sup>, Vivek B. Beechar<sup>1</sup>, Wenyi Zhu<sup>1</sup>, Yiqun Zhang<sup>3</sup>, Fengju Chen<sup>3</sup>, Gordon B. Mills<sup>4</sup>, Carrie A. Mohila<sup>5</sup>, Chad J. Creighton<sup>3,6</sup>, Jeffrey L. Noebels<sup>2,7,8</sup>, Kenneth L. Scott<sup>7,11</sup> & Benjamin Deneen<sup>1,2,9\*</sup>

Glioblastoma is a universally lethal form of brain cancer that exhibits an array of pathophysiological phenotypes, many of which are mediated by interactions with the neuronal microenvironment<sup>1,2</sup>. Recent studies have shown that increases in neuronal activity have an important role in the proliferation and progression of glioblastoma<sup>3,4</sup>. Whether there is reciprocal crosstalk between glioblastoma and neurons remains poorly defined, as the mechanisms that underlie how these tumours remodel the neuronal milieu towards increased activity are unknown. Here, using a native mouse model of glioblastoma, we develop a high-throughput in vivo screening platform and discover several driver variants of PIK3CA. We show that tumours driven by these variants have divergent molecular properties that manifest in selective initiation of brain hyperexcitability and remodelling of the synaptic constituency. Furthermore, secreted members of the glypican (GPC) family are selectively expressed in these tumours, and GPC3 drives gliomagenesis and hyperexcitability. Together, our studies illustrate the importance of functionally interrogating diverse tumour phenotypes driven by individual, yet related, variants and reveal how glioblastoma alters the neuronal microenvironment.

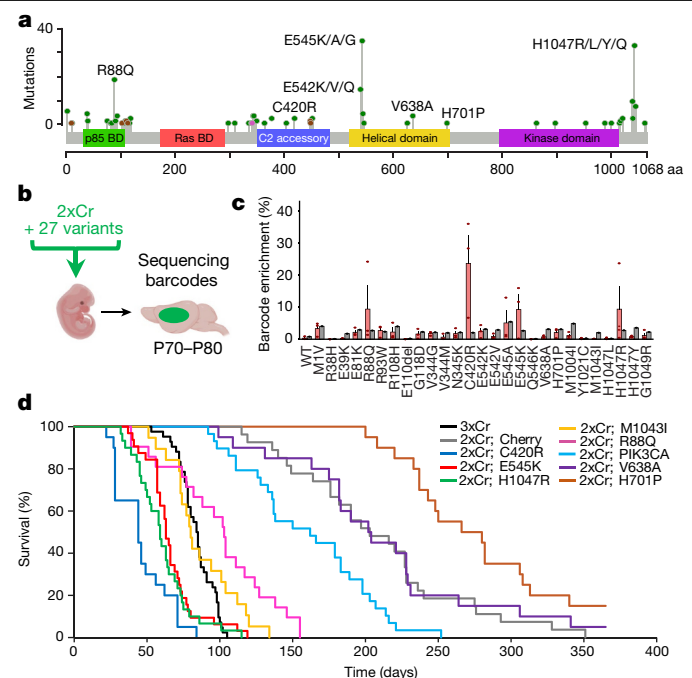
In the pursuit of cataloguing the spectrum of somatic mutations for all tumour types, cancer genomics is rapidly approaching a state of saturation mutagenesis<sup>5</sup>. These crucial advances promise to usher in a new era of personalized medicine, in which patient-specific genomic information is applied in the clinic. One major obstacle in achieving this goal is decoding driver and passenger mutations from these cohorts, as not all mutations affect malignancy, whereas others exert context-specific functions. Mutation ‘hotspots’ represent variants in key driver genes that occur with high frequency and are found across different types of cancer<sup>6</sup>. Although widely accepted as the convention for prospective selection of driver mutations, this approach overlooks most mutations and, crucially, does not account for how cellular context influences variant function<sup>7–12</sup>. Whether related variants exert differential effects on both tumour and microenvironmental phenotypes remains unknown. These limitations illustrate the need for high-throughput, functional and phenotypic screening of individual variant cohorts in contextually appropriate in vivo systems.

## Identifying PIK3CA driver variants

The RTK–RAS–PI3K pathway is a key driver of tumorigenesis across all cancers, with 90% of glioblastoma tumours exhibiting alterations in this pathway<sup>13,14</sup>. Among the specific genes in this pathway, mutations in the

PI3K catalytic subunit *PIK3CA* are found in 11% of glioblastoma tumours<sup>13</sup>. Sequencing of glioblastoma samples revealed several known hotspot mutations that drive tumorigenesis in several cancer lineages (E545K and H1047R), as well as a series of 63 in-frame mutations (as indexed in the Catalogue Of Somatic Mutations In Cancer (COSMIC)) that remain largely unclassified (Fig. 1a, Supplementary Table 1). To decode which of these PIK3CA variants function as drivers of glioblastoma, we established an in vivo complementation screening platform for glioblastoma. Our mouse glioblastoma model relies on in utero electroporation (IUE) and CRISPR–Cas9-mediated knockout of *Nf1*, *Trp53* (also known as *p53*) and *Pten* (termed ‘3xCr’), with 50% of the mice succumbing to tumours by postnatal day (P) 84<sup>15</sup> (Fig. 1d, black line, Extended Data Table 1). Removing the *Pten* guide RNA (termed ‘2xCr’) extends median survival to P203<sup>16</sup>, providing ample dynamic range to screen for factors that complement the loss of PTEN (Fig. 1d, grey line, Extended Data Table 1). Because PTEN catalyses the reverse reaction of PIK3CA—phosphorylation of phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P<sub>2</sub>, or PIP<sub>2</sub>) to phosphatidylinositol-3,4,5-trisphosphate (PtdIns(3,4,5)P<sub>3</sub>, or PIP<sub>3</sub>)—we complemented *Pten* loss with overexpression of E545K or H1047R—two bona fide hotspot driver mutations of *PIK3CA*<sup>12</sup>. Both of these variants accelerated tumour growth, which demonstrates that known PIK3CA drivers complement PTEN loss in our system (Fig. 1d, red and green lines, Extended Data Fig. 1a, b, Extended Data Table 1).

<sup>1</sup>Center for Cell and Gene Therapy, Baylor College of Medicine, Houston, TX, USA. <sup>2</sup>Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA. <sup>3</sup>Dan L. Duncan Cancer Center, Division of Biostatistics, Baylor College of Medicine, Houston, TX, USA. <sup>4</sup>Department of Cell, Developmental and Cancer Biology, Knight Cancer Institute, Oregon Health Science University, Portland, OR, USA. <sup>5</sup>Department of Pathology, Texas Children’s Hospital, Houston, TX, USA. <sup>6</sup>Department of Medicine, Baylor College of Medicine, Houston, TX, USA. <sup>7</sup>Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX, USA. <sup>8</sup>Department of Neurology, Baylor College of Medicine, Houston, TX, USA. <sup>9</sup>Department of Neurosurgery, Baylor College of Medicine, Houston, TX, USA. <sup>10</sup>These authors contributed equally: Kwanha Yu, Chia-Ching John Lin. <sup>11</sup>Deceased: Kenneth L. Scott. \*e-mail: deneen@bcm.edu

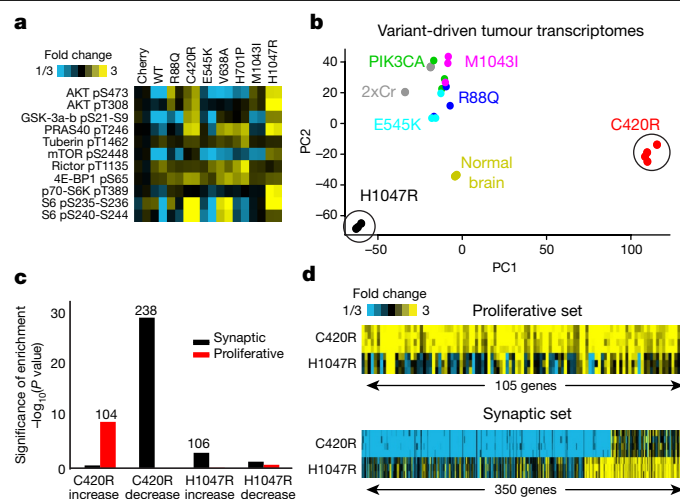


**Fig. 1 | In vivo screening identifies novel driver PIK3CA variants in glioma.** **a**, Glioblastoma-associated PIK3CA mutations indexed in COSMIC (by June 2019): missense (green), in-frame deletions and insertions (brown), or nonsense (pink) (Supplementary Table 1). **b**, Schematic of in vivo competition assay in which *Nf1* and *Trp53* are deleted (2xCr) and 27 uniquely barcoded *Pik3ca* alleles are co-electroporated. **c**, Next-generation sequencing for barcode amplification, showing the barcode for each allele (red) and the input signal (black).  $n = 3$  tumours. Data are mean and s.e.m. **d**, Kaplan–Meier curve of all tested variants.  $n = 42$  (3xCr);  $n = 28$  (Cherry);  $n = 20$  (C420R);  $n = 32$  (E545K);  $n = 30$  (H1047R);  $n = 19$  (M1043I);  $n = 21$  (R88Q);  $n = 29$  (WT);  $n = 20$  (V638A);  $n = 20$  (H701P). See Extended Data Table 1 for full survival statistics.

Our finding that established *PIK3CA* hotspots can drive in vivo tumorigenesis led us to screen a cohort of PIK3CA variants found in human glioblastoma (Fig. 1a, Supplementary Table 1). To achieve this, we used DNA barcoding together with a pooled screening strategy<sup>17,18</sup> (Methods) in our sensitized 2xCr model (Fig. 1b, Extended Data Fig. 2). After glioma formation, we used targeted sequencing of tumours for barcodes as a surrogate for variant enrichment within the tumours. This analysis demonstrated that the established hotspots (E545K and H1047R) are enriched along with other variants (R88Q and C420R) that have not been characterized in glioma (Fig. 1c). To confirm driver function of these variants, we individually expressed them in the 2xCr model, finding that they all complement the loss of PTEN and drive tumorigenesis, while also pathologically resembling glioblastoma (Fig. 1d, Extended Data Figs. 1a, b, 3, Extended Data Table 1). These data indicate that our approach can identify established hotspots, as well as new PIK3CA driver variants in glioma.

## PIK3CA tumours exhibit synaptic profiles

Although the tested driver variants accelerated tumour-associated death compared to the wild-type PIK3CA control, they did so to varying degrees (Fig. 1d). To understand a molecular basis for differences in tumorigenic potential between these variants, we performed reverse-phase protein array (RPPA) analysis on lysates collected from tumours driven by each variant, finding differential activation of PI3K pathway components (Fig. 2a). C420R and H1047R demonstrated the highest PI3K activity and were also the strongest drivers of tumorigenesis (Fig. 1d), whereas E545K, R88Q and M1043I exhibited less activation. These differences in PI3K tone and survival suggest that individual

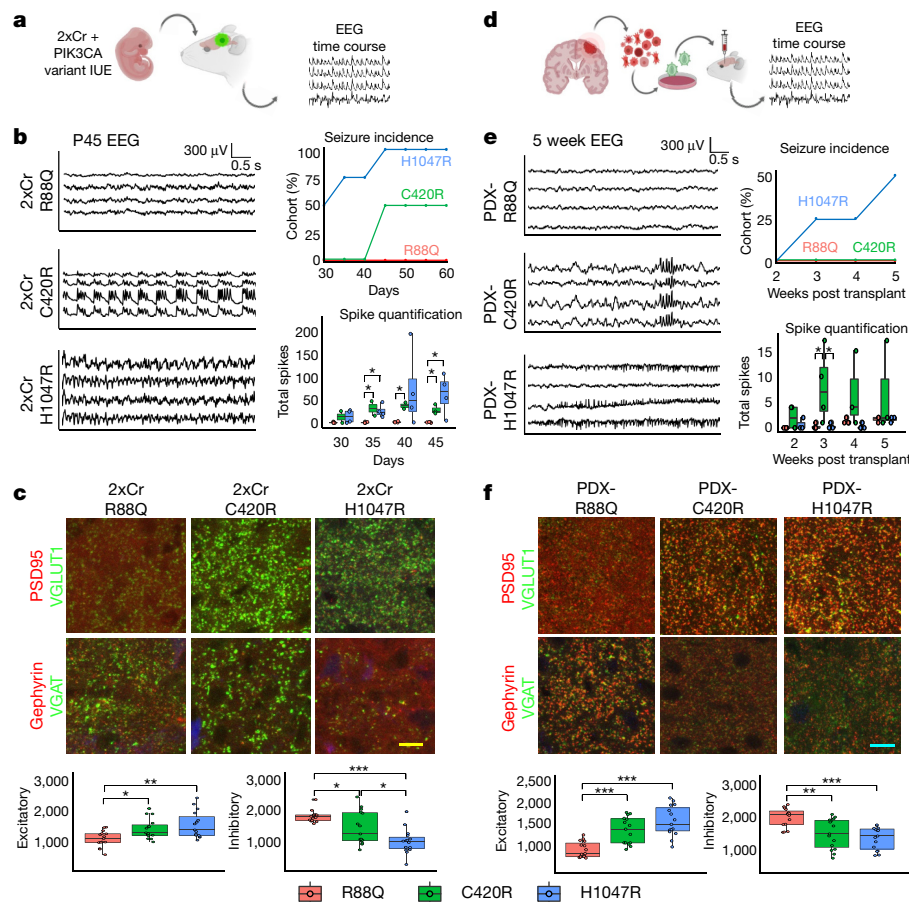


**Fig. 2 | Tumours driven by PIK3CA variants exhibit diverse molecular properties.** **a**, Tumour lysate RPPA for proteins of the PI3K–AKT–MTOR pathway<sup>12</sup>;  $n = 2$  per variant. **b**, Principal component (PC) analysis against top 2,000 variable genes across variant tumours from RNA-seq data.  $n = 2$  (normal brain);  $n = 4$  (C420R); all other variants  $n = 3$ . **c**, Enrichment analysis of the top genes increased or decreased in C420R and H1047R variants ( $P < 0.01$  for both wild-type and Cherry control comparisons, by linear model) for gene sets associated with synapse function or proliferation. Number above the bars represent genes from each set involved in observed significant overlap.  $P$  values were determined by one-sided Fisher's exact test;  $n = 4$  (C420R);  $n = 3$  (H1047R). **d**, Differential gene expression patterns (relative to average of wild-type and Cherry controls) for tumours driven by C420R or H1047R for proliferative or synaptic genes.

variants promote tumorigenesis through distinct mechanisms, which we examined via RNA sequencing (RNA-seq) analysis of end-stage tumours driven by these variants. Principal component analysis of the transcriptomes revealed that tumours driven by E545K, M1043I, R88Q, wild-type PIK3CA and the 2xCr control are highly correlated (Fig. 2b). Notably, principal component analysis revealed that the transcriptomes of tumours driven by H1047R and C420R are vastly different from this core set of variants and from one another (Fig. 2b, circles). Focusing on the molecular differences manifest in H1047R- and C420R-driven tumours, we performed Gene Ontology analysis of the transcriptome data, and found that genes associated with proliferation and synapses are differentially expressed (Fig. 2c, Supplementary Table 2). Although C420R tumours demonstrated a notable increase in proliferative genes, we found two distinct patterns of synaptic gene dysregulation, with C420R tumours showing suppression of one subgroup of synaptic genes and H1047R tumours revealing robust upregulation of an entirely different set of synaptic genes (Fig. 2d). Together, these molecular analyses reveal that gliomas driven by different PIK3CA variants are endowed with distinct molecular features that can be uncoupled from PI3K activation (that is, C420R and H1047R). Moreover, these findings highlight how driver variants endowed with single amino acid differences, exhibiting seemingly similar phenotypic effects, can have distinct underlying molecular properties.

## Selective induction of hyperexcitability

Given the distinct molecular features of tumours driven by C420R and H1047R mutations, we next determined whether these tumours exhibit unique biological properties. Because tumours driven by C420R display a pronounced increase in genes associated with proliferation (Fig. 2c, d) and cause a shorter median survival time in mice when compared to H1047R-driven tumours (44 versus 59 days), we evaluated tumour growth and proliferation during early tumorigenesis. Using BrdU labelling and



**Fig. 3 | C420R and H1047R tumours promote hyperexcitability and synaptic imbalance across tumour models.** **a**, Schematic of experimental plan for IUE model. **b**, EEG traces of select variant tumour mice at P45 and quantification of EEG analyses. Traces for C420R and H1047R occurred during seizures. Seizure incidence (top) and quantification of spike activity (bottom) are shown;  $n = 4$  mice per variant. Data in seizure incidence graph denote mean values. Box plots denote the median value, interquartile range (25th–75th percentiles), with whiskers extending to 1.5 times the interquartile range. **c**, Antibody staining of excitatory and inhibitory synapses in P30 mouse brains at peritumoral margins (top) and quantification from a 34,000- $\mu\text{m}^2$  field of view (bottom). GFP signal (blue, not labelled) denotes tumour.  $n = 3$  mice;  $n = 5$

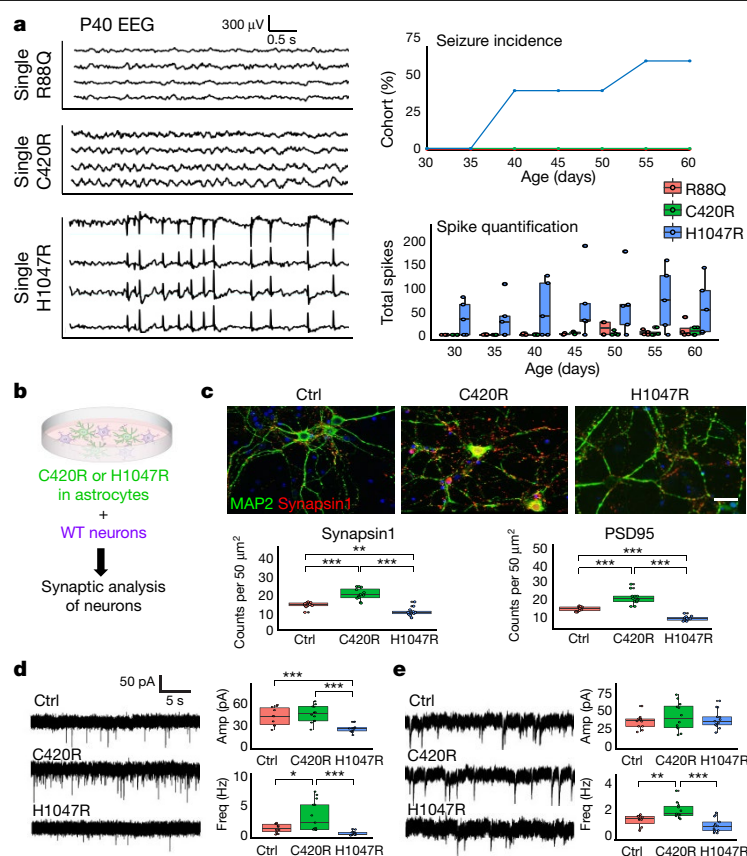
technical replicates. Box plots as in **b**. **d**, Schematic of PDX model experiments. **e**, EEG traces (left) and quantification (right) of mice bearing PDX tumours expressing PIK3CA variants five weeks after transplantation. Trace for PDX-C420R demonstrates aberrant EEG patterns without associated seizures; trace for PDX-H1047R was during an electrographic, non-convulsive seizure.  $n = 4$  mice per variant. Box plots are as in **b**. **f**, Antibody staining of excitatory and inhibitory synapses from peritumoral margins of PDX brains (top) and quantification from a 34,000- $\mu\text{m}^2$  field of view (bottom).  $n = 3$  mice;  $n = 5$  technical repeats. Box plots as in **b**. Scale bars, 12.5  $\mu\text{m}$  (c) and 10  $\mu\text{m}$  (f). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , one-way analysis of variance (ANOVA).

MRI imaging (Extended Data Fig. 1c–e), we found that C420R-driven tumours exhibit an enhanced proliferative phenotype compared to tumours driven by H1047R and R88Q, thus corroborating the underlying molecular features of these variants. The other defining molecular feature of C420R- and H1047R-driven tumours is dysregulation of distinct cohorts of synapse-associated genes (Fig. 2c, d). It is well-established that synaptic imbalance can result in extensive changes in network excitability, which in some cases culminates in seizure activity<sup>19,20</sup>. Crucially, seizures are an early pathophysiological hallmark of malignant glioma that are recapitulated in our mouse model of glioma<sup>15,21–23</sup>. Therefore, we determined whether mice bearing tumours driven by C420R and H1047R exhibit pronounced network hyperexcitability during the early stages of tumour progression (Fig. 3a). We performed serial video electroencephalography (EEG) recordings every 5 days during the P30–P60 interval (Extended Data Fig. 4), and found that mice bearing tumours driven by H1047R show an earlier onset of seizures (50% of the cohort exhibited convulsive seizures at P30) than mice bearing tumours driven by C420R (seizures first appear at P45) (Fig. 3b). EEG recordings of these mice during this interval revealed considerable increases in network hyperexcitability, with both groups of mice demonstrating similar increases

in spiking frequency as early as P30–P35 (Fig. 3b). By contrast, mice bearing R88Q-driven tumours did not demonstrate convulsive seizures or changes in EEG activity during the P30–P60 interval (Fig. 3b). These variant-specific increases in hyperexcitability correlate with the relative changes in synaptic gene cohorts of C420R and H1047R tumours, which suggests that tumours driven by these variants result in early synaptic imbalance during the formative stages of tumour progression.

To determine whether C420R and H1047R tumours exhibit changes in synaptic phenotypes in the tumour microenvironment during early tumorigenesis, we isolated C420R- and H1047R-driven tumours from P30 mice and stained for makers of excitatory (VGLUT1 and PSD95) and inhibitory (VGAT and gephyrin) synapses at the peritumoral margins (Extended Data Fig. 5). These studies revealed significant increases in excitatory synapses at the peritumoral margins of both C420R and H1047R tumours compared with R88Q tumours (Fig. 3c). Notably, analysis of inhibitory synapses revealed a decrease in their numbers surrounding C420R and H1047R tumours, compared to R88Q (Fig. 3c). These changes in synaptic constituency—increases in excitatory and decreases in inhibitory synapses—in neurons adjacent to C420R- and H1047R-driven tumours further indicate that the observed increases





**Fig. 4 | C420R and H1047R differentially promote synaptic imbalance.** **a**, EEG traces and quantification of variants overexpressed in mouse brains at P40. Trace for H1047R was during seizure. Seizure incidence and quantification of total spike activity are shown;  $n = 4$  mice per variant. Box plots are as in Fig. 3b. No significant ( $P < 0.05$ )  $P$  values were calculated by one-way ANOVA. **b**, Schematic of astrocyte–neuron co-culture. WT, wild type. **c**, Top, antibody staining for synapsin1 (red) and MAP2 (green) of astrocyte–neuron co-cultures; astrocytes infected with control (ctrl) or C420R or H1047R viruses.

Bottom, quantification of PSD95 and synapsin1 staining.  $n = 12$  technical replicates per condition. Data are mean and s.e.m. Scale bar, 50  $\mu$ m. **d**, **e**, Representative traces (left) and quantification (right) of excitatory (**d**) or inhibitory (**e**) postsynaptic currents from whole-cell recording of neurons co-cultured on astrocytes infected with control, or C420R or H1047R viruses.  $n = 12$  technical replicates per condition. Data are mean and s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , one-tailed independent  $t$ -test; Tukey's test was used to compare individual mean values.

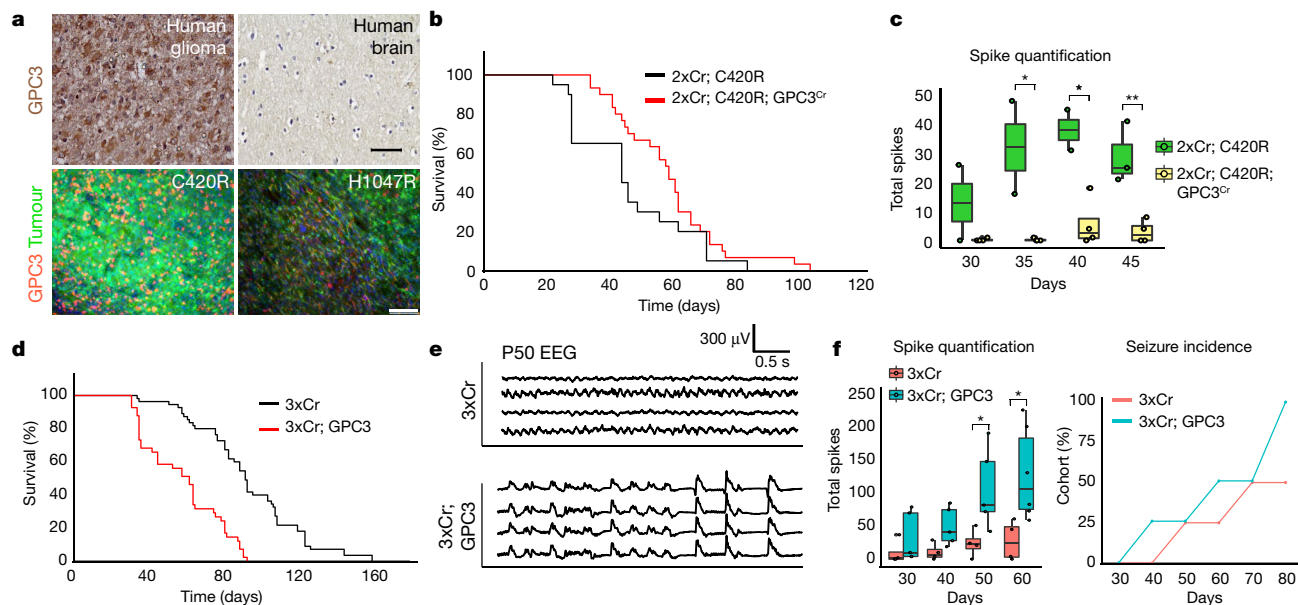
in network hyperexcitability result, in part, from synaptic imbalance. Next, we extended these functional studies to patient-derived xenograft models (PDX), by generating glioma stem-cell lines<sup>24</sup> that individually overexpress H1047R, C420R and R88Q (Fig. 3d). These lines were subsequently transplanted into the brains of SCID mice, in which we assessed the onset of hyperexcitability via EEG, and found that mice bearing PDX-H1047R exhibited early onset of seizures whereas mice bearing PDX-C420R exhibited early onset of network hyperexcitability (Fig. 3e). PDX-R88Q did not demonstrate early onset of either seizures or hyperexcitability (Fig. 3e). Moreover, cellular analysis of the peritumoral synaptic constituency revealed that both PDX-H1047R and PDX-C420R tumours provoke an increase in excitatory synapses, coupled with a decrease in inhibitory synapses compared to PDX-R88Q tumours (Fig. 3f). These alterations in synaptic constituency are met with congruent changes in cell proliferation, as PDX-H1047R and PDX-C420R exhibit increased incorporation of BrdU (Extended Data Fig. 1f). Together, these functional studies in PDX models reinforce our conclusion that these variants selectively promote changes in network hyperexcitability and synaptic constituency.

### Variant-specific synaptogenic mechanisms

The synaptic changes at the peritumoral margin in both the native and PDX models suggest that these variants function via

cell-non-autonomous mechanisms. However, several PIK3CA variants are associated with monogenic paediatric overgrowth syndromes and can promote hyperexcitability on their own, including H1047R, which suggests that they may also function cell autonomously<sup>25–27</sup>. To address this, we performed individual IUE, outside the tumorigenic context, with H1047R, C420R and R88Q, in conjunction with serial EEG recordings every 5 days during the P30–P60 interval (Extended Data Fig. 4). Mice expressing H1047R exhibited increased hyperexcitability as early as P30, coupled with seizures at P40, whereas mice expressing C420R and R88Q did not demonstrate any changes during this period (Fig. 4a); we did not observe any changes in cell proliferation (Extended Data Fig. 1g–i). These data indicate that H1047R promotes hyperexcitability on its own, whereas C420R cannot—which suggests that these variants use different mechanisms to alter synaptic constituency. Next, we used an established neuron–glial co-culture system<sup>15</sup>, which enables us to directly assess the cell-non-autonomous contributions of these variants to synaptogenesis (Fig. 4b). We overexpressed C420R and H1047R in mouse astrocytes co-cultured with neurons, and found that C420R promotes synapse formation between neurons as measured by the expression of synaptic markers (Fig. 4c). Whole-cell recordings revealed increased frequency of excitatory and inhibitory postsynaptic currents (EPSCs and IPSCs, respectively) in neurons when co-cultured with astrocytes expressing C420R (Fig. 4d, e). By contrast, overexpression of H1047R in astrocytes did not enhance synapse formation on neurons





**Fig. 5 | GPC3 promotes gliomagenesis and synaptic imbalance.** **a**, Top, GPC3 staining of human glioblastoma and healthy human brain control. Staining was not independently repeated. Bottom, GPC3 antibody staining of C420R and H1047R tumours from P30 mouse brains.  $n = 4$  mice for each variant. These experiments were independently repeated four times. Scale bars, 50  $\mu\text{m}$  (top) and 100  $\mu\text{m}$  (bottom). **b**, Kaplan–Meier curve comparing GPC3 deletion (GPC3<sup>cr</sup>) in 2xCr; C420R tumours.  $n = 20$  (C420R) and  $n = 30$  (C420R; GPC3<sup>cr</sup>) mice. Statistics are in Extended Data Fig. 8b. **c**, Quantification of total spike activity shows that EEG spike activity decreases with GPC3 deletion.  $n = 4$  mice

per variant. Box plots are as in Fig. 3b. \* $P < 0.05$ , \*\* $P < 0.01$ , one-way ANOVA.  $n = 4$  mice each. **d**, Kaplan–Meier curve comparing *Gpc3* overexpression in 3xCr tumours.  $n = 55$  (3xCr) and  $n = 41$  (3xCr; GPC3) mice. Statistics are in Extended Data Fig. 8b. **e**, EEG traces of *Gpc3* overexpression in 3xCr tumours at P50. **f**, Quantification of total spike activity (left) and seizure incidence (right).  $n = 4$  mice (3xCr) and  $n = 5$  mice (3xCr; GPC3) for spike quantifications;  $n = 4$  mice per condition for seizure incidence. Box plots are as in Fig. 3b. \* $P < 0.05$ , one-way ANOVA.

(Fig. 4d, e). Together, these results illustrate mechanistic differences between H1047R and C420R. H1047R promotes brain hyperexcitability on its own, through cell-autonomous mechanisms, whereas C420R promotes brain hyperexcitability specifically in the context of a glial tumour, through cell-non-autonomous mechanisms.

### GPC3 promotes tumorigenesis

To understand how C420R exerts a cell-non-autonomous effect on synaptogenesis, we further analysed the RNA-seq data from the variant-driven tumours, focusing on a set of genes that is secreted from glia and promotes synapse formation onto neurons<sup>28</sup>. Among this group, we found that members of the glypican family are selectively upregulated in tumours driven by C420R (Extended Data Fig. 6a). Analysis of The Cancer Genome Atlas (TCGA) data for glioma revealed that glypican 3 (*GPC3*) is more highly expressed in human glioblastoma tumours than in lower-grade glioma (Extended Data Fig. 6b). We confirmed protein expression using immunohistochemistry on primary human glioma samples and detected increased expression of GPC3 in C420R-driven mouse tumours (Fig. 5a, Extended Data Fig. 6c). Together, these data suggest that GPC3 may contribute to glioma tumorigenesis and downstream synaptic remodelling. Notably, how secreted synaptogenic proteins expressed in glioma influence (1) tumorigenesis, (2) the surrounding neuronal microenvironment and (3) key neurological features of glioma pathophysiology remain largely undefined. Therefore, examining the function of GPC3 in this context provides an opportunity to understand how these facets of glioma biology are regulated.

To determine whether GPC3 contributes to C420R-driven tumours, we used CRISPR–Cas9 approaches to delete GPC3 (GPC3<sup>cr</sup>) in C420R-driven mouse tumours, and found that loss of GPC3 extended median survival (Fig. 5b, P44 versus P60, Extended Data Fig. 7). Notably, serial EEG recordings of mice bearing 2xCr; C420R; GPC3<sup>cr</sup> tumours revealed a loss of early onset hyperexcitability compared to C420R-driven controls

(Fig. 5c). Having established that GPC3 is necessary for key pathophysiology properties of C420R-driven tumours, we next evaluated whether it is sufficient to drive glioma tumorigenesis independently of PIK3CA variants. We combined PiggyBac-mediated *Gpc3* overexpression with our 3xCr system; survival analysis revealed that mice overexpressing *Gpc3* exhibited significantly shorter median survival than 3xCr controls (P59 versus P91) (Fig. 5d, Extended Data Fig. 7). Next, we examined seizure phenotypes via video EEG in 3xCr tumours overexpressing *Gpc3*, and found that mice bearing *Gpc3*-overexpressing tumours exhibited early onset of convulsive seizures and heightened network hyperexcitability as early as P40 (Fig. 5d–f). Together, these studies indicate GPC3 functions as a driver of glioma tumorigenesis that accelerates the onset of seizures and hyperexcitability.

The robust effects of *Gpc3* overexpression on these key tumour pathophysiological properties led us to examine the cellular phenotypes in these tumours. First, we evaluated cellular proliferation, and found an increase in BrdU labelling in GPC3-driven tumours (Extended Data Fig. 8a). Next, we probed the peritumoral synaptic constituency, and identified increases in both excitatory and inhibitory synapses in these tumours (Extended Data Fig. 8b), which suggests that GPC3 promotes the aberrant formation of both excitatory and inhibitory synapses. Previous studies have demonstrated that other members of the glypican family promote synapse formation onto neurons through secretion from adjacent glial cells<sup>29,30</sup>, suggesting that GPC3 may act by a similar mechanism. To test this, we overexpressed *Gpc3* in mouse astrocytes co-cultured with neurons and found that *Gpc3* overexpression in astrocytes markedly increases synapse formation onto neurons, as measured by expression of synaptic markers and whole-cell recordings from these neurons (Extended Data Fig. 9a–d). Conditioned medium from these astrocytes promoted heightened spontaneous postsynaptic current frequency on neurons (Extended Data Fig. 9e–g), which indicates that GPC3 acts via non-cell-autonomous mechanisms. These studies identify GPC3 as a key factor in

glioma-induced neosynaptogenesis that manipulates the neuronal microenvironment during tumour progression.

## Discussion

Decoding variant function in cancer has become an increasingly crucial requirement for developing personalized therapeutic agents. In this study, we describe an *in vivo* complementation screening platform that enables the rapid identification of driver mutations in a native model of glioma. Using this system, we identified several driver variants of PIK3CA that are active in glioma. Crucially, a subset of these variants (C420R, M1043I and R88Q) appear to function in a glioma-relevant manner, as studies in other cancer models did not identify these variants, or their relative oncogenic potential varies<sup>7,17</sup>. Together, these studies illustrate that variant function is context-specific and highlight the importance of decoding cell lineage and genetic relationships<sup>7,31–33</sup>. Furthermore, among the PIK3CA variants that operate as drivers in our glioma system, transcriptome analysis identified marked molecular differences in the tumours driven by these variants. These findings demonstrate that similar variants can engender a diverse range of molecular properties in glioma tumours, reinforcing the importance of deciphering the selective functions of each driver variant.

Our identification of proliferative and synaptic gene dysregulation among tumours driven by H1047R and C420R served as an entry point for demonstrating phenotypic differences among these tumours. Using both native mouse and PDX models, we found that tumours driven by these variants selectively alter the neuronal microenvironment towards heightened network hyperexcitability and seizure onset—two key features of glioma pathophysiology<sup>22</sup>. Furthermore, H1047R and C420R use different mechanisms to engender these common physiological phenotypes. Nevertheless, these findings further illustrate that crosstalk between glioma tumours and the surrounding neuronal microenvironment is a key contributor to malignant growth. Several recent studies have shown that increases in neuronal activity promotes glioma tumour proliferation<sup>1,2</sup>, and when put together with our findings, indicate a vicious cycle between glioma cells and adjacent neurons, in which glioma cells promote network hyperexcitability and these increases in neuronal activity promote tumour growth. To understand how these variant-driven tumours promote neuronal activity, we identified GPC3 as a driver of glioma tumorigenesis, network hyperexcitability and synapse formation. Because GPC3 promotes synapse formation through cell-non-autonomous mechanisms, a model emerges in which secreted proteins from glioma stimulates neuronal activity by triggering synaptogenesis during early tumorigenesis, working together with neurons to orchestrate both network hyperexcitability and growth.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-1952-2>.

1. Venkatesh, H. S. et al. Targeting neuronal activity-regulated neuroligin-3 dependency in high-grade glioma. *Nature* **549**, 533–537 (2017).
2. Venkatesh, H. S. et al. Neuronal activity promotes glioma growth through neuroligin-3 secretion. *Cell* **161**, 803–816 (2015).
3. Venkatesh, H. S. et al. Electrical and synaptic integration of glioma into neural circuits. *Nature* **573**, 539–545 (2019).
4. Venkataramani, V. et al. Glutamatergic synaptic input to glioma cells drives brain tumour progression. *Nature* **573**, 532–538 (2019).
5. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
6. Samuels, Y. et al. High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**, 554 (2004).
7. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
8. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
9. Miller, M. L. et al. Pan-cancer analysis of mutation hotspots in protein domains. *Cell Syst.* **1**, 197–209 (2015).
10. Ng, P. K. et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33**, 450–462 (2018).
11. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
12. Zhang, Y. et al. A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations. *Cancer Cell* **31**, 820–832 (2017).
13. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
14. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
15. John Lin, C. C. et al. Identification of diverse astrocyte populations and their malignant analogs. *Nat. Neurosci.* **20**, 396–405 (2017).
16. Kfoury, N. et al. Cooperative p16 and p21 action protects female astrocytes from transformation. *Acta Neuropathol. Commun.* **6**, 12 (2018).
17. Dogruluk, T. et al. Identification of variant-specific functions of PIK3CA by rapid phenotyping of rare mutations. *Cancer Res.* **75**, 5341–5354 (2015).
18. Tsang, Y. H. et al. Functional annotation of rare gene aberration drivers of pancreatic cancer. *Nat. Commun.* **7**, 10500 (2016).
19. Nelson, S. B. & Valakh, V. Excitatory/inhibitory balance and circuit homeostasis in autism spectrum disorders. *Neuron* **87**, 684–698 (2015).
20. Ramocki, M. B. & Zoghbi, H. Y. Failure of neuronal homeostasis results in common neuropsychiatric phenotypes. *Nature* **455**, 912–918 (2008).
21. Huberfeld, G. & Vecht, C. J. Seizures and gliomas—towards a single therapeutic approach. *Nat. Rev. Neurol.* **12**, 204–216 (2016).
22. van Breemen, M. S., Wilms, E. B. & Vecht, C. J. Epilepsy in patients with brain tumours: epidemiology, mechanisms, and management. *Lancet Neurol.* **6**, 421–430 (2007).
23. Campbell, S. L. et al. GABAergic disinhibition and impaired KCC2 cotransporter activity underlie tumor-associated epilepsy. *Glia* **63**, 23–36 (2015).
24. Lang, F. M. et al. Mesenchymal stem cells as natural biofactories for exosomes carrying miR-124a in the treatment of gliomas. *Neuro-oncol.* **20**, 380–390 (2018).
25. Keppler-Noreuil, K. M. et al. PIK3CA-related overgrowth spectrum (PROS): diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. *Am. J. Med. Genet. A.* **167A**, 287–295 (2015).
26. Mirzaa, G. et al. PIK3CA-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. *JCI Insight* **1**, 87623 (2016).
27. Roy, A. et al. Mouse models of human PIK3CA-related brain overgrowth have acutely treatable epilepsy. *eLife* **4**, e12703 (2015).
28. Allen, N. J. & Eroglu, C. Cell biology of astrocyte–synapse interactions. *Neuron* **96**, 697–708 (2017).
29. Allen, N. J. et al. Astrocyte glypicans 4 and 6 promote formation of excitatory synapses via GluA1 AMPA receptors. *Nature* **486**, 410–414 (2012).
30. de Wit, J. et al. Unbiased discovery of glypican as a receptor for LRRTM4 in regulating excitatory synapse development. *Neuron* **79**, 696–711 (2013).
31. Shah, M. A., Denton, E. L., Arrowsmith, C. H., Lupien, M. & Schapira, M. A global assessment of cancer genomic alterations in epigenetic mechanisms. *Epigenetics Chromatin* **7**, 29 (2014).
32. Thomas, R. K. et al. High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.* **39**, 347–351 (2007).
33. Zhang, J., Wu, L. Y., Zhang, X. S. & Zhang, S. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* **15**, 271 (2014).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

# Article

## Methods

All mouse gliomas were generated in the CD-1 IGS mouse background as previously described<sup>15</sup>. IUE was performed on embryonic day 15. Previously generated CRISPR constructs were used to knockout *Nf1*, *Pten* and *Trp53* (each at 1.5 µg per µl)<sup>15</sup>. *Pik3ca* alleles were genomically integrated and overexpressed through the piggyBac transposase (PBase) system. The pGlast-PBase plasmid (at 2.0 µg per µl)<sup>34</sup> was co-electroporated with a PBCAG-PIK3CA\* construct (1.0 µg per µl) (see below). In addition, the PBCAG-GFPt2aLuc (1.0 µg per µl) was co-electroporated to allow for fluorescent and bioluminescent visualization.

In vivo functionalization studies of *Gpc3* were performed by co-electroporating a PBCAG-GPC3 construct (at 1.0 µg per µl), generated through the HiTMMob approach (see below), into the 3xCr model<sup>15</sup>. Loss-of-function studies were performed by co-electroporating a CRISPR construct targeting the third exon of the mouse *Gpc3* gene. The gRNA sequence (5'-CTTGGGTCTGATCAACG-3') was generated through Broad Institute GPP portal (<https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design>)<sup>35</sup>. Candidate gRNA sequence was validated through the mismatch-cleavage SURVEYOR assay (IDT, 706020) on genomic DNA (gDNA) obtained from tumour tissue (Extended Data Fig. 7c). Primer sequences for SURVEYOR are as follows: ON (ATACCAGCAATGAATGTGAGTCAA and CCACCCACC ACCACAATGAAG), OT1 (TTGCACAGCATAGCCTGAGA and TACCCTCTA CTGCTGACCCC), OT2 (GGCATCTTGAGTGTCTCTACCA and TTCTGAA GGAGAACTATGCTACC), OT4 (TGGAAGTAAGGAACAGGCCCC and TGGCTCCCACTCTAACTCCTT) and OT5 (AAGCATCCGTCGCTACCATC and AAAGCTTGGCTGGAACCTGT). All procedures were approved by the Institutional Animal Care and Use Committee (IACUC) at Baylor College of Medicine and conform to the US Public Health Service Policy on Human Care and Use of Laboratory Animals.

### In vivo barcode enrichment competition assay

A piggyBac transposable vector (PBCAG-EGFP-T2A-GWR1R4) was engineered from the PBCAG-EGFP construct<sup>34</sup>. The eGFP STOP codon was removed and an inframe T2A sequence followed by attR1/attR4 Gateway cloning sites were inserted. These attR sites flanked chloramphenicol and ccdB selection cassettes. A V5 tag sequence was also inserted downstream of the attR4 site. *Pik3ca* alleles and barcode sequences were cloned in using the HiTMMob approach<sup>17,18</sup>. The list of tested PIK3CA variants and associated mutational information are given in Supplementary Table 1. PIK3CA variant constructs were generously provided in collaboration with G.B.M., K.L.S. and the CTD<sup>2</sup> (<https://ocg.cancer.gov/programs/ctd2>).

Pooled injection cocktails were assembled such that the pool of tested variants totalled 1.0 µg per µl rather than a single PIK3CA variant. In brief, equal moles of each plasmid were mixed together, ethanol precipitated and re-dissolved/concentrated in ddH<sub>2</sub>O. This pooled plasmid mixture was diluted to 1.0 µg per µl in the final IUE injection cocktail.

After IUE of the pooled cocktail (with 2xCr, PBase and GFPt2aLuc), mice were born and observed for symptoms suggestive of brain tumours (see below). After demonstration of symptoms, the mouse was euthanized and tumours dissected with the aid of the fluorescence reporter. Subsequently, gDNA was prepped with the EZNA Tissue DNA Kit (Omega Bio-tek, D3396), according to the manufacturer's instructions. Samples were prepared in biological and technical replicates ( $n = 3$ –5 biological replicates;  $n = 3$  technical replicates). In addition, the IUE injection cocktail was used for input, prepared in technical duplicates.

After gDNA isolation, barcoded libraries were prepared as previously reported<sup>17,18</sup>. PCR reactions amplified the barcoded pools from 50 ng gDNA (experimental samples) or 2 ng of plasmid pool (input control) using Platinum Super Mix (ThermoFisher, 12532016) with primers targeting the T3 promoter site (directly upstream of the BC) (5'-CAATTAACCCCTCACTAAAGG-3') and the V5 tag

(5'-ACCGAGGAGAGGGTTAGGGAT-3'). Amplification parameters were as follows: 1× (94 °C 4 min); 35× (94 °C 1 min, 54 °C 1 min, 68 °C 1 min); 1× (68 °C 10 min); 10 °C hold. PCR products were purified with the PureLink PCR Purification Kit (ThermoFisher, K310001), processed using the Ion Plus Library Kit (ThermoFisher, 4471252), subsequently purified and ligated to unique Ion Xpress Barcode Adaptors (ThermoFisher, 4474517). The resulting Ion Xpress barcoded libraries were amplified, purified, and pooled for PGM sequencing (318 V2 Chip) following the manufacturer's recommendations. Raw data were concatenated into one 'reference' file and indexed using Burrows–Wheeler alignment tool for alignment of barcode sequences (with parameters '-l7 -t12 -N -n3') for counting the occurrence of each barcode. Barcode enrichment was assessed by quantifying the number of occurrences for each barcode sequence as a ratio to total number of barcode reads in each sample. Standard error of the mean was calculated across replicates and plotted as error bars on the barcode enrichment graphs.

### Bioluminescence imaging

D-Luciferin (150 µg; PerkinElmer, 122799) per gram body weight was delivered through intraperitoneal injection. Ten minutes after injection, bioluminescence signal was acquired for 5 min, followed by a 10-s X-ray exposure for skeletal imaging. Bioluminescence scans were initiated at P60 and performed every 30 days thereafter.

### Brain tumour collection for histology

Unless a time point was specified, tumour-bearing mice were observed for symptoms suggestive of tumours including—but not limited to—lethargy, hunched posture, decreased appetite, decreased grooming, trembling/shaking, squinting eyes, partial limb paralysis and abnormal gait, denoting the IACUC permitted endpoint. After demonstration of symptoms, mice were humanely euthanized and fixed through intracardial perfusion of 4% paraformaldehyde in PBS. After perfusion and dissection, the brain was further fixed overnight in fresh 4% paraformaldehyde solution overnight and then preserved in 70% ethanol for eventual paraffin embedding. For frozen preservation, after perfusion and dissection, brains were sunk in 20% sucrose in PBS overnight and frozen into Tissue-Tek OCT Compound (Sakura Finetek, 4583) and stored at -80 °C until cryostat sectioning.

### Histological analysis

For haematoxylin and eosin (H&E) staining, 10-µm paraffin-embedded sections were processed as follows: 3 × 3 min in xylene, 3 × 3 min in 100% ethanol, 3 × 3 min in 95% ethanol, 3 min in 80% ethanol, 5 min in 70% ethanol, 5 min in ddH<sub>2</sub>O, 2.5 min in Harris haematoxylin (Poly Scientific R&D, S212A), running tap water wash, 30 s in 95% ethanol, 2.5 min in eosin (Poly Scientific R&D, S176), 2 × 2 min in 95% ethanol, 2 × 2 min in 100% ethanol, and 2 × 2 min in xylene. Staining was preserved with Permount Mounting Media (Electron Microscope Sciences, 17986-01) under a coverslip. Histological diagnoses of mouse-IUE-generated tumours were validated across  $n \geq 6$  tumours per variant.

For immunohistochemistry, frozen brains were sectioned to 20–40 µm thickness. Sections were subject to antigen retrieval (when needed), blocking and primary antibody incubation overnight at 4 °C. The following primary antibodies were used: rat anti-BrdU (BU1/75 (ICR1), 1:200; abcam, ab6326), mouse anti-gephyrin (1:500; Synaptic Systems, 147011), rabbit anti-GFP (1:1,000; ThermoFisher, A-11122), goat anti-GPC3 (W-18, 1:100; Santa Cruz Biotechnology, SC-10455), rabbit anti-human HLA A (EP1395Y, 1:100, ABCAM, ab52922), mouse anti-PSD95 (7E3-1B8, 1:500; ThermoFisher, MA1-046), guinea-pig anti-VGAT (1:500; Synaptic Systems, 131004), guinea-pig anti-VGLUT1 (1:2,000; Millipore, AB5905). We used species-specific secondary antibodies tagged with Alexa Fluor 488, 568, or 647 (1:1,000, ThermoFisher) for immunofluorescence. After Hoechst nuclear counter staining (ThermoFisher, H3570, 1:50,000), coverslips were mounted with VECTASHIELD antifade mounting medium (Vector Laboratories, H-1000). DAB

horseradish peroxidase (HRP) (Vector Laboratories, SK-4100) was used for chemical colorimetric detection following species-specific, HRP-conjugated secondary antibody labelling.

To assay *in vivo* cell proliferation, 4 h before collecting, 100 µg BrdU (in PBS) per gram body mass was delivered through intraperitoneal injection. Mouse brains were collected, frozen and sectioned as described above. Before blocking, sections were incubated in 2 N HCl at 37 °C for 30 min and neutralized with 3.8% sodium borate for 10 min at room temperature.

Excitatory and inhibitory synapses were quantified with the Synapse Counter plugin for ImageJ<sup>36</sup>. The images were taken with the Leica SP8 confocal microscope at 0.31-µm intervals over a 5-µm depth (15 optical sections) across replicates ( $n = 5$  technical replicates;  $n = 3$  biological replicates).

The human tumour tissue arrays (US Biomax, T175a) contained a spectrum of brain tumours that included high-grade gliomas, meningiomas and T cell lymphoma. Non-disease brain sections were also present in the array. Sections were rehydrated similar to paraffin-embedded sections above, stained and dehydrated-preserved similar to our H&E staining protocol.

### RPPA

GFP-positive tumour samples were homogenized with lysis buffer (1% Triton X-100, 50 mM HEPES, pH 7.4, 150 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 1 mM EGTA, 100 mM NaF, 10 mM Na pyrophosphate, 1 mM Na<sub>3</sub>VO<sub>4</sub>, 10% glycerol, protease inhibitor (Roche, 05056489001), and phosphatase inhibitors (Roche, 04906837001) on ice. Cellular debris was removed by centrifugation (4 °C, 14,000 r.p.m., 10 min). The lysates were mixed with 4× SDS buffer and diluted to a final protein concentration of 1–1.5 µg µl<sup>-1</sup>. Samples were prepared in biological duplicates ( $n = 2$ ). Samples were probed with 287 antibodies through the Functional Proteomics RPPA Core Facility at MD Anderson Cancer Center. The full list of tested antibodies, source and working dilution are available at: <https://www.mdanderson.org/research/research-resources/core-facilities/functional-proteomics-rppa-core/antibody-information-and-protocols.html><sup>10</sup>.

### RNA-seq analysis

Similar to descriptions above, tissue samples were collected with the aid of fluorescence microscopy from end-stage tumours. Total RNA was isolated using the RNeasy Plus Mini Kit (Qiagen, 74134) according to the manufacturer's protocol. Samples were prepared in biological replicates,  $n \geq 3$  per variant genotype. RNA integrity (RNA integrity number  $\geq 8.0$ ) was confirmed using the High Sensitivity RNA Analysis kit (AATI, DNF-472-0500) on a 12-Capillary Fragment Analyzer. Illumina sequencing libraries with 6-bp single indices were constructed from 1 µg total RNA using the TruSeq Stranded mRNA LT kit (Illumina, RS-122-2101). The resulting library was validated using the Standard Sensitivity NGS Fragment Analysis Kit (AATI, DNF-473-0500) on a 12-Capillary Fragment Analyzer. Equal concentrations (2 nM) of libraries were pooled and subjected to sequencing of approximately 20 million reads per sample using the Mid Output v2 kit (Illumina, FC-404-2001) on a Illumina NextSeq550 following the manufacturer's instructions.

### Bioinformatics

Sequencing alignment and transcript abundance estimation (fragments per kilobase of transcript per million mapped reads, or FPKM) was performed using HISAT and Cufflinks. Combat software<sup>37</sup> was used to alleviate any potential batch effects due to substantial time between sequencing runs. In the downstream RNA-seq analysis, we compared each experimental (variant) group with the wild-type group and with the Cherry control group (two separate comparisons) by contrasts using linear models on log-transformed (base 2) expression values, in which all samples are used to estimate the variance of each gene when using the full model (linear model) in design, providing greater

statistical power. For each experimental group, genes significant versus both control and wild-type groups were defined, using  $P < 0.01$  and fold change in the same direction for both comparisons. Expression heat maps were generated using JavaTreeView<sup>38</sup>.

### MRI

Mice were imaged at the Small Animal MR Imaging Core at Baylor College of Medicine using a 9.4 T horizontal bore magnet with a 60-mm inner diameter microgradient and a 35-mm inner diameter radiofrequency volume resonator (Bruker BioSpin MRI GmbH) once a week starting at 4 weeks of age. Once anaesthetized (induction at 3–4% isoflurane, maintenance at 2–3% in 100% oxygen at 3 l min<sup>-1</sup>), mice were placed on an animal bed, head first, in the prone position and inserted into the bore of the magnet. Normal body temperature was maintained during the imaging sessions. The respiration rate of the mice (typical range, 30–40 breaths min<sup>-1</sup>) was monitored throughout the entire experiment with an abdominal pneumatic pillow (SA Instruments).

For *in vivo* brain imaging of each mouse, three low-resolution scans (axial, sagittal and coronal slice orientations) and one high-resolution scan (axial orientation) were acquired in succession in a single imaging session. T2-weighted images were obtained using a fast spin echo pulse sequence with excitation and refocusing flip angles of 90° and 180° respectively at a rare factor of 8. In each low-resolution brain scan (repetition time/echo time = 2,500/33 ms), 10–12 slices per brain volume were scanned in a field-of-view of 4.0 × 4.0 cm<sup>2</sup>, with a matrix size of 256 × 256 pixels yielding a spatial resolution of 0.156 × 0.156 mm<sup>2</sup> with a slice thickness of 1.0 mm and an interslice distance of 1.5 mm, taking 2 averages per slice with a scan time per brain volume of 2 min 40 s. Each high-resolution brain scan (repetition time/echo time = 2,500/36 ms) was taken with 14–18 slices per brain volume, scanned in a field-of-view of 3.5 × 3.5 cm<sup>2</sup>, with a matrix size of 384 × 384 pixels yielding a spatial resolution of 0.091 × 0.091 mm<sup>2</sup> with a slice thickness of 0.75 mm and an interslice distance of 1.0 mm, taking 10 averages per slice with a scan time per brain volume of 8 min.

Relative tumour area was calculated as the product of the height and width of the tumour along the dorsal–ventral and medial–lateral axis, respectively. The plane (along the rostral–caudal axis) that contained the largest tumour section at the last scan was used for analysis.

### In vivo video EEG recording

Similar to previous descriptions<sup>15</sup>, IUE mice at 3 weeks of age were anaesthetized by isoflurane vaporization pump (induction at 3–4%, maintenance at 2–3% in 100% oxygen at 1 l min<sup>-1</sup>) and surgically implanted with a bilateral silver wire electrode (0.005-inch diameter) attached to a microminiature connector. Electrodes were placed in the left and right, frontal and parietal regions. EEG and behavioural activity in freely moving mice were analysed using simultaneous video-EEG monitoring (Haramonie software version 6.1c, Stellate systems). All EEG signals were filtered using a 60-Hz notch filter, 0.3-Hz high-pass filter and 70-Hz low-pass filter. Mouse EEG monitoring was initiated at 30 days of age, and continued at 5 or 10 day intervals with a 24-h recording period per session. Seizure activity was quantified by visual inspection of the EEG waveform and corresponding video-recorded behaviour. The frequency of interictal spike activity was calculated from EEG waveform data of each 24-h recording period using Matlab scripts. At least  $n = 4$  mice were analysed for each condition.

### Astrocyte–neuron co-culture and immunocytochemistry

Cortical astrocyte cultures were prepared from P1–P3 newborn wild-type mice. The cortex was dissected, removing the meninges, and coarsely chopped with surgical scissors, followed by enzymatic dissociation of papain supplemented with DNase I (Worthington Biochemical, LK003150) at 37 °C for 15 min. Enzymatic dissociation was neutralized with 10% fetal bovine serum (FBS; ThermoFisher, 16000044) in DMEM/F12 (ThermoFisher, 11320033). After two centrifugation and

# Article

PBS washes, cells were suspended in astrocyte culture medium (DMEM/F12, 10% FBS, 1% penicillin/streptomycin; ThermoFisher, 15140122), filtered through a cell strainer, and plated onto a poly-D-lysine-coated T75 flask. Once confluent, flasks were mechanically agitated to enrich for astrocytes.

Astrocytes were virally infected selected through antibiotics. After selection, approximately  $8 \times 10^4$  astrocytes were seeded on poly-D-lysine-coated 12 mm coverslips in a 24-well culture dish. Around 24 h later, approximately  $6 \times 10^4$  cortical neurons were seeded on the top of the astrocytes and maintained in neuronal culture medium (Neurobasal medium (ThermoFisher, 21103049) supplemented with B27 (ThermoFisher, 17504044), gentamicin (ThermoFisher, 15750060) and GlutaMAX (ThermoFisher, 35050061)).

For immunocytochemistry, similar to previous descriptions<sup>15</sup>, co-cultures were washed and fixed with 4% paraformaldehyde for 15 min, then rinsed with PBS, blocked and stained with primary antibodies overnight at 4 °C. The following antibodies were used: chicken anti-MAP2 (1:1,000; EnCor Biotech, CPCA-MAP2), rabbit anti-synapsin-I (1:1,000; Millipore, AB1543) and mouse anti-PSD95 (K28/43, 1:1,000; UC Davis/NIH NeuroMab Facility, 75-028). Similar to the above immunohistochemistry, species-specific Alexa Fluor 488 or 568 were used for secondary antibodies, followed by nuclear counter-staining and coverslip mounting.

## Electrophysiology

Whole-cell recordings were performed on neurons co-cultured with astrocytes (after viral infection and selection) in parallel at days in vitro (DIV) 7 and 14, similar to a previous description<sup>15</sup>. Neurons were recorded in a chamber with continuous flow of external solution at a fixed flow rate controlled by a flow valve. For each recording, data were filtered with a 60-Hz notch filter and analysed with template-based event detection algorithms in Clampfit 10.6 (Molecular Devices).

Standard artificial cerebral spinal fluid (140 mM NaCl, 2.4 mM KCl, 10 mM HEPES, 10 mM glucose, 4 mM MgCl<sub>2</sub>, 2 mM CaCl<sub>2</sub>; pH 7.3; osmolarity approximately 300 mOsm) was used as the external solution. The internal solution contained 136 mM KCl, 17.8 mM HEPES, 1 mM EGTA, 0.6 mM MgCl<sub>2</sub>, 4 mM ATP, 0.3 mM GTP, 12 mM creatine phosphate, and 50 U per ml phosphocreatine kinase. The recording pipettes used in the experiment had a resistance of 3–6 MΩ. Antagonists (3 mM kynurenic acid and 20 μM bicuculline) were applied sequentially with 3 min of washing in between. All chemicals were purchased from Sigma-Aldrich.

To determine the glutamatergic or GABAergic identities of the spontaneous activities recorded, the spontaneous synaptic potential was first recorded for 2 min for baseline activities, as well as activities under the treatment of 3 mM kynurenic acid and 20 μM bicuculline (2 min for each treatment with 3 min of no treatment in between application of the two antagonists). An event with a rise time of approximately 0.5 ms and decay time of less than 20 ms were considered to be a glutamatergic event, whereas an event with a rise time of around 0.5 ms and a decay time of more than 20 ms were considered to be a GABAergic response.

For the GPC3 condition medium study, as previously described<sup>29</sup>, COS-7 cells (ATCC, CRL-1651) cultured in neuronal culture medium (see above) were transfected with PBCAG-GPC3 (PBCAG-GFP as control) using Lipofectamine 2000 Transfection Reagent (ThermoFisher, 11668019) according to manufacturer's instructions. Conditioned medium was collected and concentrated with Spin-X UF concentrator columns (Corning, CLS431482). After quantification of protein concentration using a standard Bradford assay, protein was diluted into neuronal culture medium to 80 μg per ml, and this medium was added onto cortical neuron cultures at DIV1. Whole-cell recordings and immunocytochemistry were performed at DIV 9. Whole-cell recordings were done with any antagonists to measure total spontaneous postsynaptic potential. Neurons co-cultured with wild-type primary astrocytes were used as a positive control.

## PDX model

Patient-derived primary glioblastoma cell lines were maintained in neurosphere medium (DMEM/F12, ThermoFisher, 11320082), supplemented with B27 (1X, ThermoFisher, 17504001), bFGF (20 ng ml<sup>-1</sup>, Peprotech, 100-18B) and EGF (20 ng ml<sup>-1</sup>, Peprotech, 100-47). Cells were lentivirally infected with different *PIK3CA* mutants (or a mCherry control reporter). After 48 h of infection, cells were selected for by puromycin (1 μg ml<sup>-1</sup>, ThermoFisher, A1113803). Approximately 50,000 live cells were stereotactically transplanted into the brains of 6-week-old male ICR-SCID (Taconic, ICRSC-M) mouse brains (from bregma, cells were injected 1.0 mm lateral, 2.0 mm caudal, 2.5 mm ventral). Similar to previously described approaches<sup>39</sup>, a week after transplant, EEG recording electrodes were implanted and recorded from once a week for the subsequent four weeks.

## Statistical analyses

The log-rank test was used to compare survival differences across groups for Kaplan–Meier survival analysis, and values are listed in Extended Data Table 1. One-way ANOVA was used to compare BrdU proliferation (across technical replicates), MRI tumour growth, EEG spike quantification and synapse quantification differences between group means, followed by Tukey's test to compare individual means. Independent *t*-test was used to compare differences across groups for EPSC and IPSC frequency and amplitude quantifications. Significant differences are denoted by asterisks in associated graphs, and an absence of asterisks denotes no significant difference unless otherwise stated. Distribution of data was assumed to be normal, but this was not formally tested. Randomization of animal studies was used in the data analysis. No data points were excluded from analyses. Mice were excluded from analyses if they did not demonstrate reporter (GFP) activity after IUE, signifying unsuccessful electroporation. No statistical methods were used to predetermine sample size, and investigators were not blinded to allocation during experiments and outcome assessment.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The RNA-seq data of tumours driven by *PIK3CA* variants have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE123519. All other data in this article are available from the corresponding author upon reasonable request.

## Code availability

No custom code was used. R package limma eBayes function was used to define differentially expressed genes. Bioconductor SVA/Combat package was used for batch correction.

34. Chen, F. & LoTurco, J. A method for stable transgenesis of radial glia lineage in rat neocortex by piggyBac mediated transposition. *J. Neurosci. Methods* **207**, 172–180 (2012).
35. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
36. Dzyubenko, E., Rozenberg, A., Hermann, D. M. & Faissner, A. Colocalization of synapse marker proteins evaluated by STED-microscopy reveals patterns of neuronal synapse distribution in vitro. *J. Neurosci. Methods* **273**, 149–159 (2016).
37. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
38. Saldanha, A. J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
39. Buckingham, S. C. et al. Glutamate release by primary brain tumors induces epileptic activity. *Nat. Med.* **17**, 1269–1274 (2011).

**Acknowledgements** This study is dedicated to the memory of our dear friend and colleague, Kenneth L. Scott, as his intellect, enthusiasm and collaborative spirit were a driving force in



this endeavour. This work was supported by grants from the Cancer Prevention Research Institute of Texas (RP150334 and RP160192 to B.D., K.L.S., C.A.M. and C.C.), National Cancer Institute-Cancer Therapeutic Discovery (U01-CA217842 to B.D., G.B.M. and K.L.S.), National Institutes of Health (R01-CA223388 to B.D. and J.L.N.; R01-NS071153 to B.D.; T32-HL902332 to K.Y.), the American Cancer Society-Rob Rutherford Glioblastoma Research Postdoctoral Fellowship (PF-15-220-01-TBG to K.Y.), and Howard Hughes Medical Institute Gilliam Fellowship (A.H.). We acknowledge the assistance of the Baylor College of Medicine Mouse Phenotyping Core with funding from the NIH (U54-HG006348). This project was supported by the BCM Small Animal MRI and Texas Children's Hospital Small Animal Imaging Facility. Functional Proteomics RPPA Core Facility at MD Anderson Cancer Center, this facility is funded by NCI CA16672. We thank F. F. Lang for providing patient-derived cell lines under the auspices of his Internal Review Board protocol (LAB04-001) post-de-identification.

**Author contributions** K.Y., C.-C.J.L., J.L.N., K.L.S. and B.D. designed the experiments, and interpreted results; K.Y. established the PIK3CA screening platform and generated all the IUE- and PDX-PIK3CA tumour-bearing mice; K.K. and K.Y. generated the barcoded PIK3CA libraries; K.Y., B.L., V.B.B. and G.B.M. performed the RNA-seq and RPPA analysis; C.J.C., Y.Z. and F.C. performed the all bioinformatics analysis; A.H. performed all the EEG studies, with assistance

from K.Y.; J.L.N. assisted in interpretation of EEG studies; K.Y. and C.-C.J.L. performed all the synaptic staining; C.-C.J.L. performed the co-culture studies and whole-cell recordings; C.A.M. provided neuropathological support; E.H.H. generated the GSC lines with variant overexpression; C.-C.J.L. performed the GPC3 experiments; W.Z. and Y.-T.C. generated reagents for GPC3 experiments. K.Y., J.L.N. and B.D. wrote the manuscript. B.D. and K.L.S. conceived the project; B.D. supervised all aspects of this work; K.Y. and C.-C.J.L. contributed equally.

**Competing interests** The authors declare no competing interests.

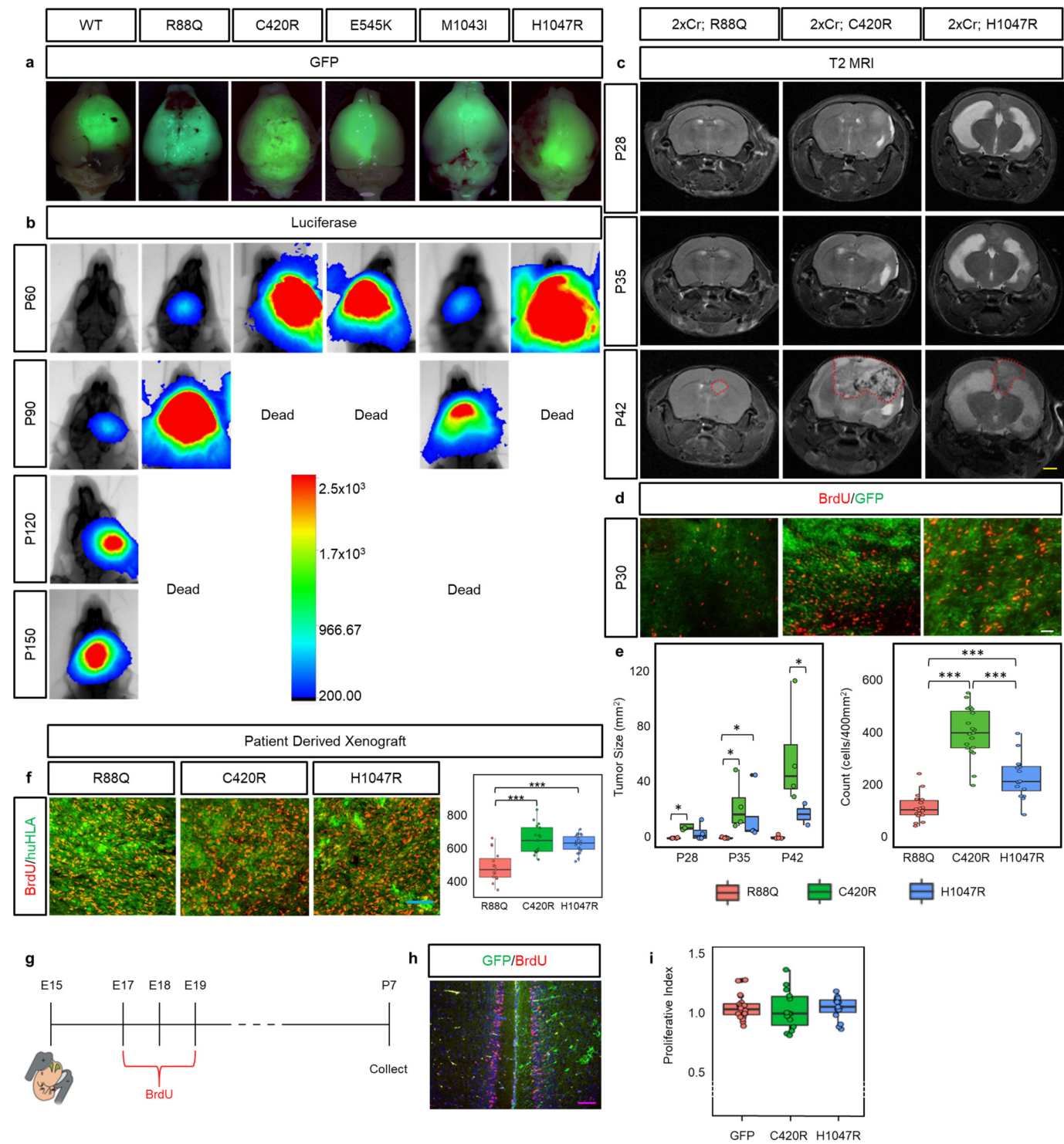
**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-1952-2>.

**Correspondence and requests for materials** should be addressed to B.D.

**Peer review information** *Nature* thanks Yuan Zhu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

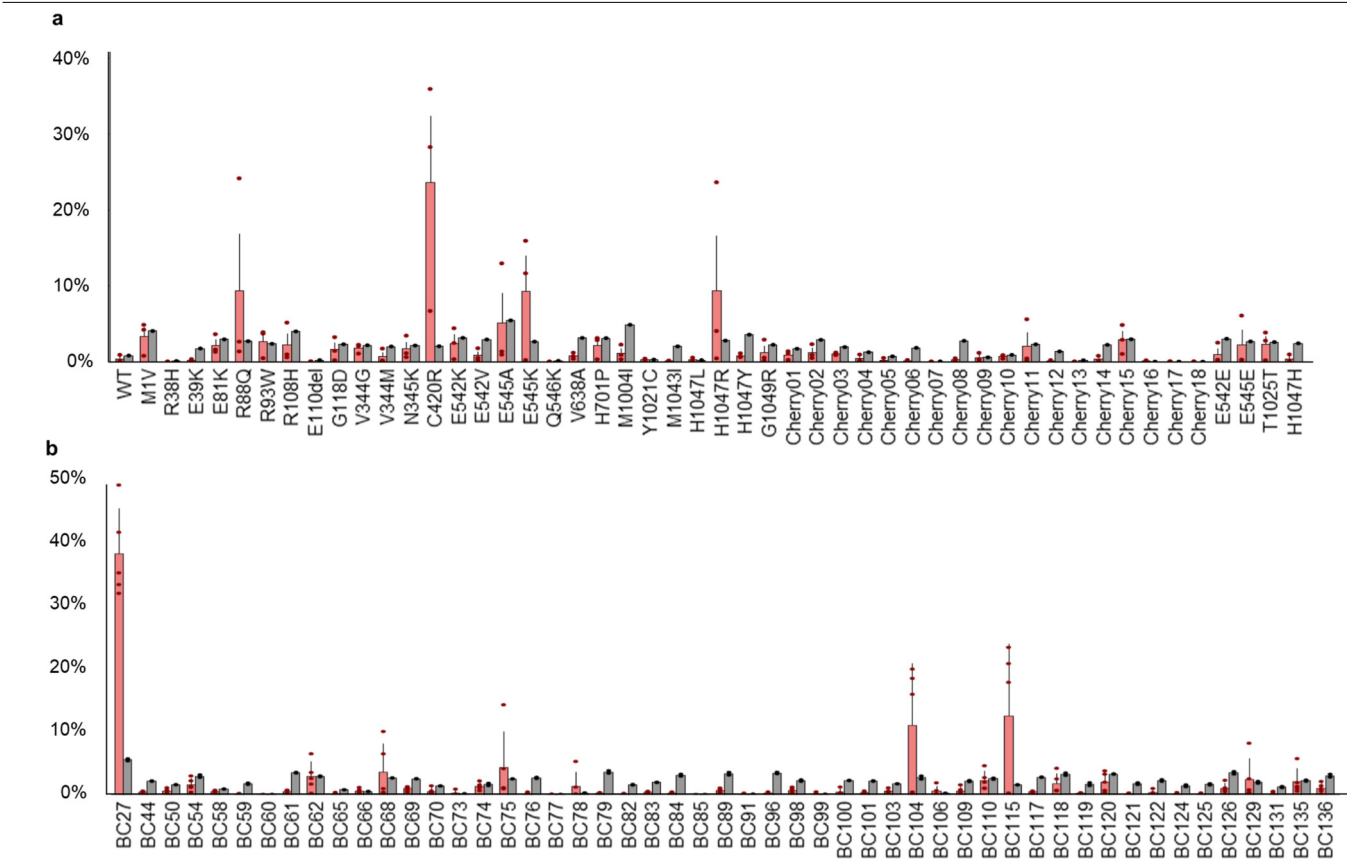
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1|See next page for caption.

**Extended Data Fig. 1 | Differential tumour growth across PIK3CA variant tumours.** **a**, Whole-brain GFP fluorescence image overlaid on a brightfield image representing characteristic tumours at time of death. **b**, Representative bioluminescence intensity images taken at 1-month intervals starting at 2 months of age, reflective of median survival trend. Scale bar for bioluminescence intensity quantifies photon counts over the 5 min of IVIS recording. Experiments were independently repeated three times with similar results for each variant. **c**, Longitudinal T2 MRI of variant tumours. Readings were taken at 1-week intervals from 4 weeks of age. All images are spatially matched along the rostral–caudal axis. Red dotted outline at P42 denotes the tumour boundary. Yellow scale bar, 2.5 mm.  $n = 4$  mice for each variant. These experiments were not independently repeated. **d**, BrdU (red) antibody staining on 30-day-old mouse brains. White scale bar, 40  $\mu\text{m}$ .  $n = 4$  mice for each variant.  $n = 18$  (R88Q), 19 (C420R) and 16 (H1047R) technical repeats. **e**, Associated quantification for relative tumour area from MRI and BrdU incorporation

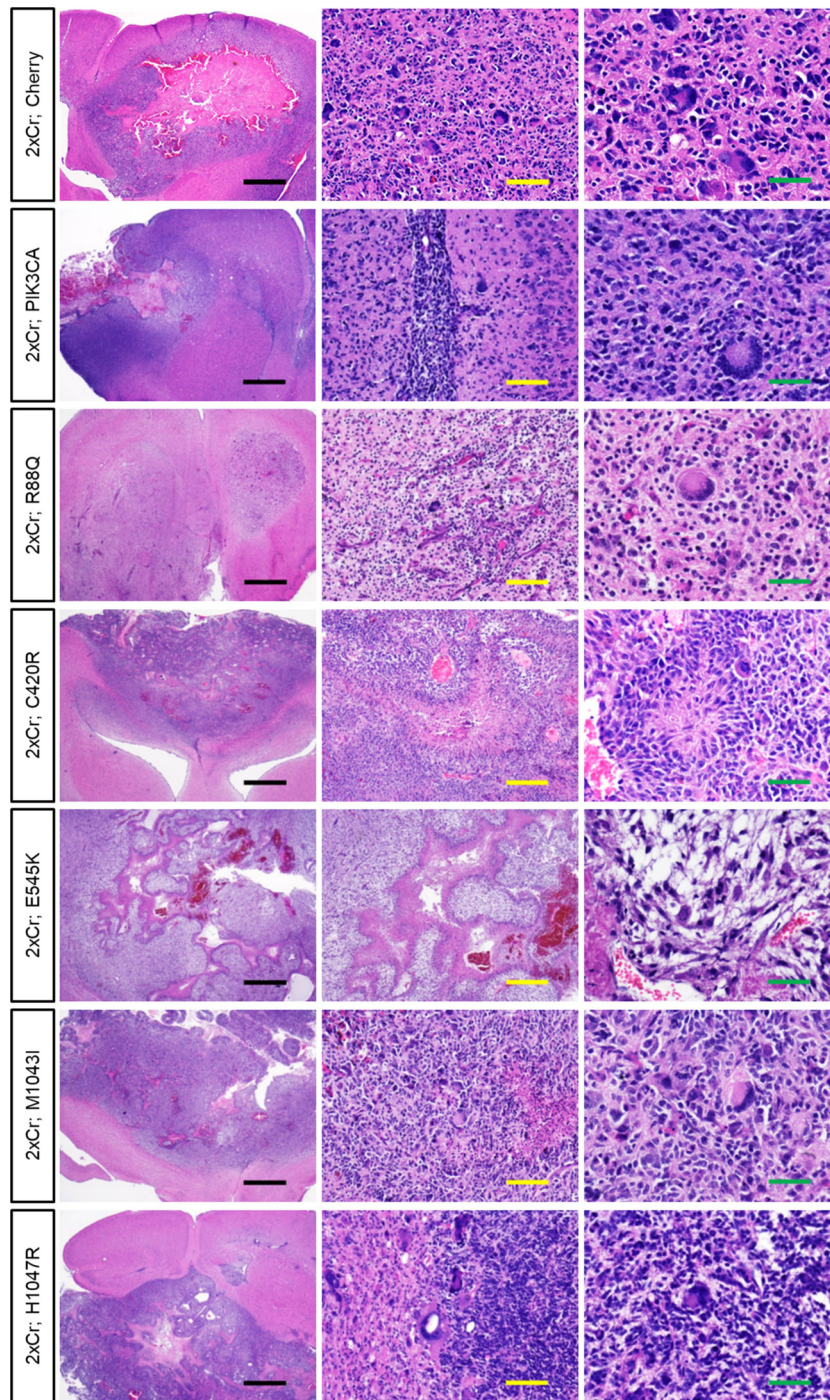
analysis. Box plots are as in Fig. 3b.  $^*P < 0.05$ ,  $^{***}P < 0.001$ , one-way ANOVA. **f**, Left, immunofluorescence analysis of BrdU (red) incorporation on PDX tumour sections. Human tissue was identified by staining for human HLA (green). Blue scale bar, 100  $\mu\text{m}$ . Right, quantification of BrdU incorporation in a 2,000  $\mu\text{m}^2$  area.  $n = 4$  mice;  $n = 5$  technical repeats. Box plots are as in Fig. 3b.  $^{***}P < 0.001$ , one-way ANOVA. **g**, Schematic illustrating experimental approach and timeline. **h**, Representative image of control GFP IUE. Proliferative index was calculated by dividing the number of BrdU<sup>+</sup> cells on the electroporated side (marked by GFP) by the number of BrdU<sup>+</sup> cells on the non-electroporated, contralateral side. Purple scale bar, 100  $\mu\text{m}$ . **i**, Quantification of proliferative index for activating variants C420R and H1047R along with control (GFP), demonstrating no significant difference in proliferation.  $n = 3$  mice;  $n = 5$  technical repeats. Box plots are as in Fig. 3b.  $P$  values were not significant ( $P > 0.05$ , one-way ANOVA).



**Extended Data Fig. 2 | In vivo competition assay identified PIK3CA driver variants.** **a**, Results of next-generation barcode sequencing from 2xCr tumour tissue co-electroporated with all tested alleles pooled together. Pool includes listed variants along with wild-type (WT), 4 silent mutants (E542E, E545E, T1025T and H1047H) and 18 uniquely barcoded Cherry constructs as controls.  $n = 3$  tumours. Data are mean and s.e.m. **b**, Barcode sequencing for 2xCr

tumours co-electroporated with the H1047R allele (tagged with barcode sequence 27), demonstrating single amplification when diluted with 50 different passenger barcodes (uniquely barcoded Cherry constructs).  $n = 2$  tumours with 2–3 replicates; red bars denote tumour samples; grey bars denote input. Data are mean and s.e.m.

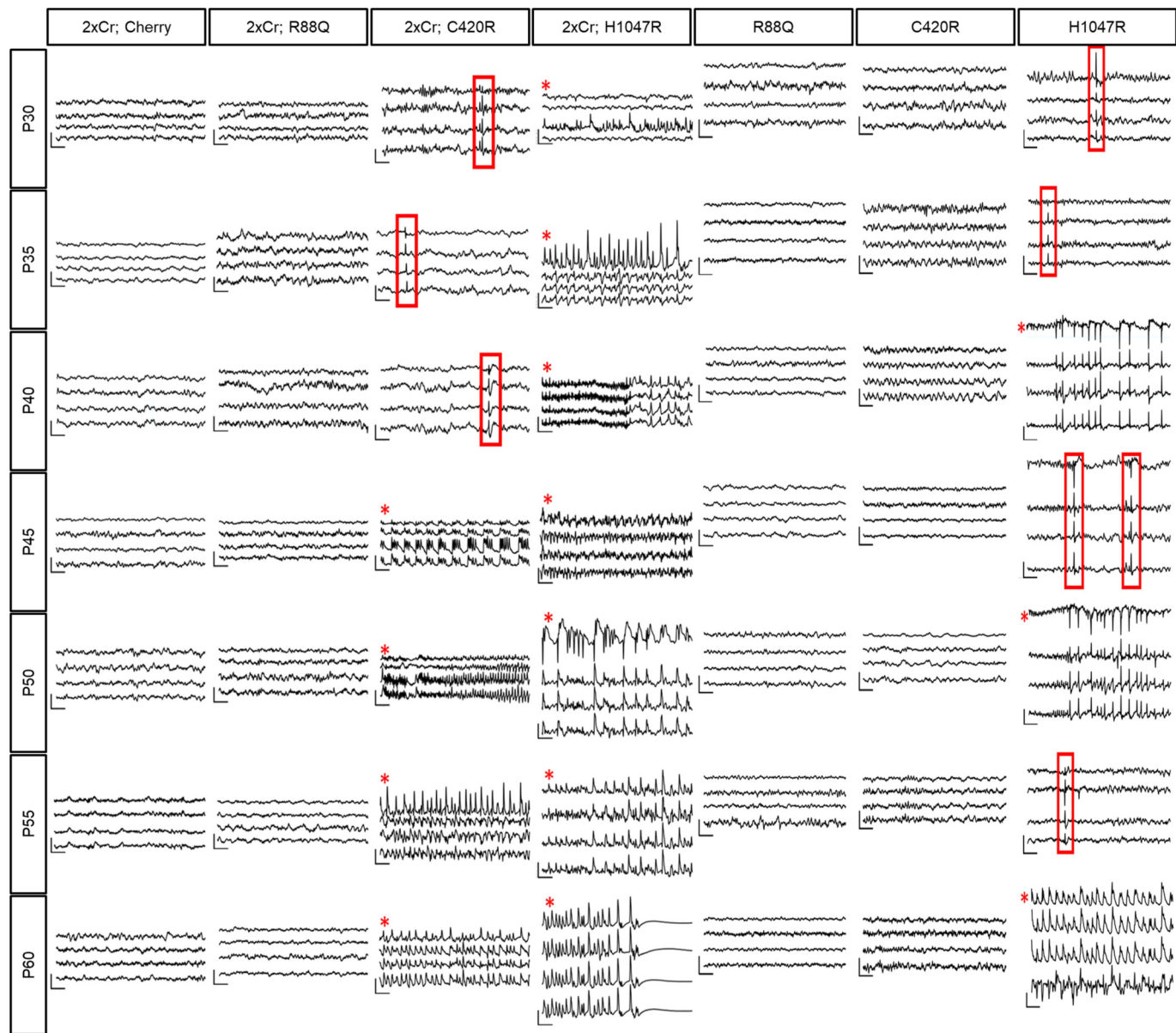




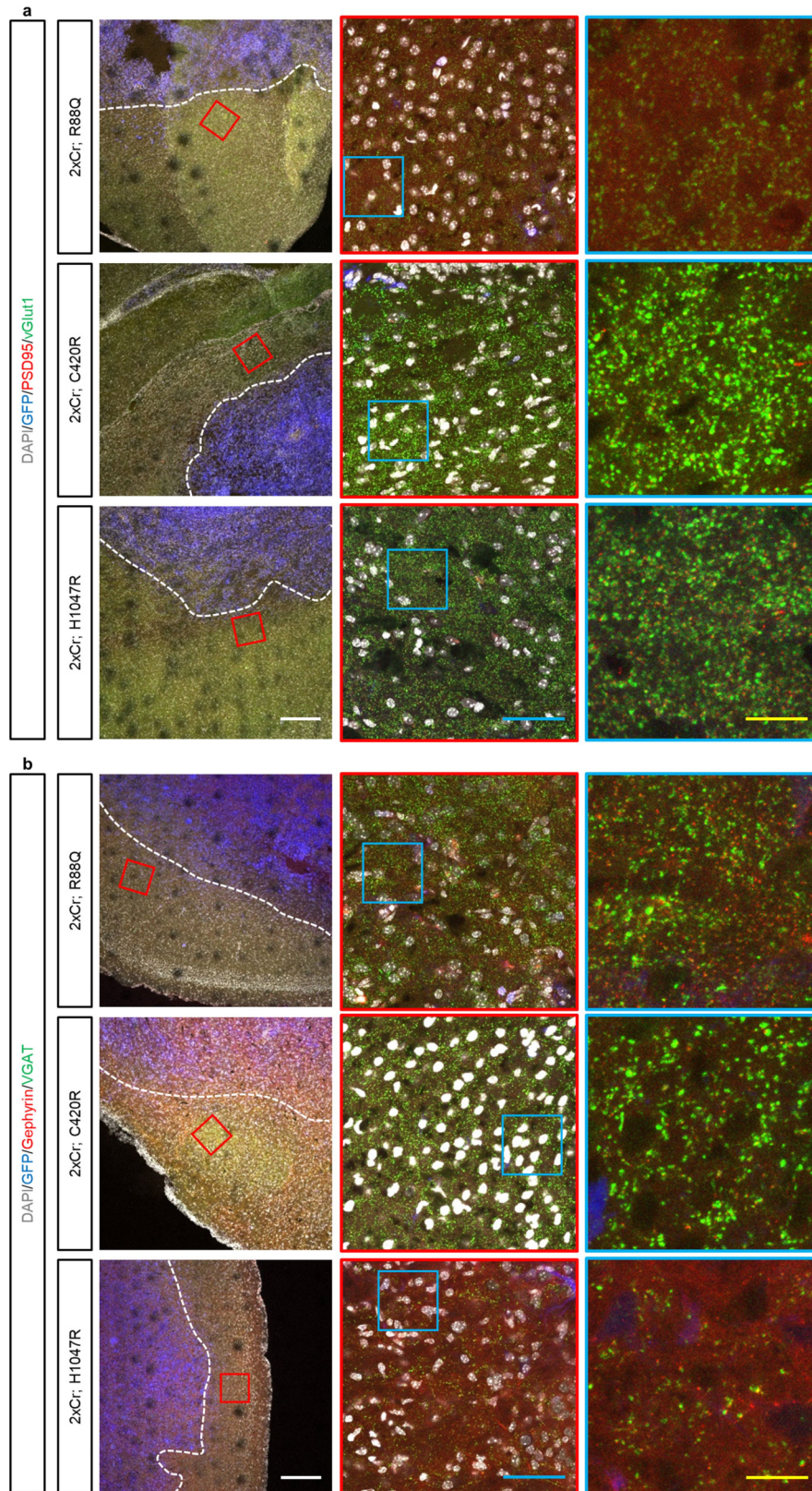
**Extended Data Fig. 3 | PIK3CA variants did not alter tumour histopathology.** H&E staining of brains containing hypercellular and infiltrative high-grade gliomas with pleomorphic tumour cells. All tumours are histologically graded as high-grade glioma, either WHO (World Health Organization) grade III

anaplastic astrocytoma or grade IV glioblastoma. Black scale bars, 1 mm; yellow scale bars, 100  $\mu$ m; green scale bars, 50  $\mu$ m. Representative images of each variant-driven tumour are from  $n = 6$  brains.





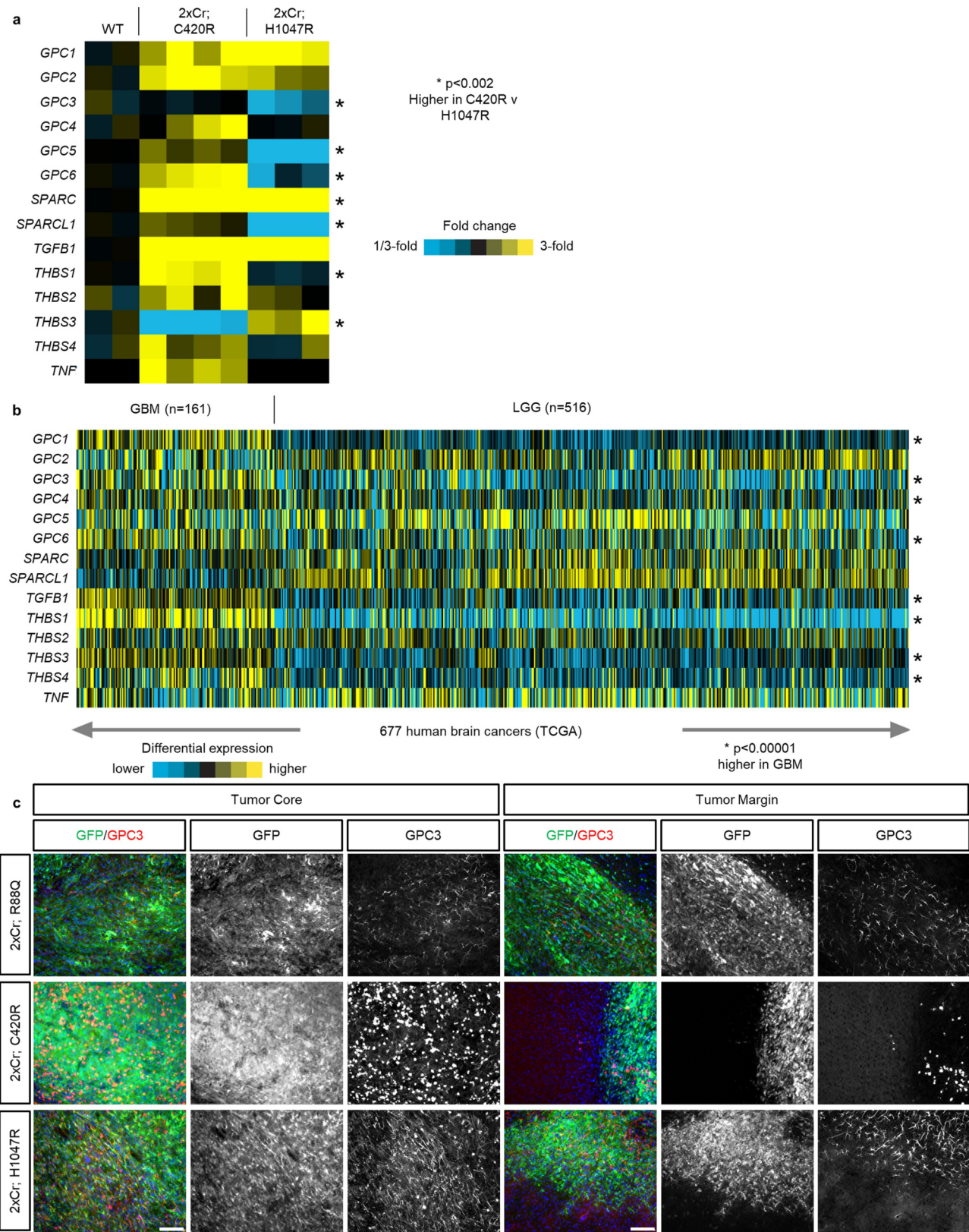
**Extended Data Fig. 4 | PIK3CA variant differentially promote both tumour-associated and unassociated seizures.** Longitudinal EEG recordings from PIK3CA variant tumour brain (2xCr) and non-tumour brains, starting at P30. Red boxes outline interictal spike activity. Red asterisks denote generalized seizures confirmed with simultaneous videos. Traces plot from top to bottom are recordings from the left frontal, left parietal, right frontal and right parietal brain regions. Traces are representative of four mice per variant. Vertical scale bars, 300  $\mu$ V; horizontal scale bars, 0.5 s.



**Extended Data Fig. 5 | PIK3CA variants differentially alter the local synaptic constituency at the peritumoral margins. a, b**, Immunofluorescence analysis of tumour brains, stained for excitatory (a) and inhibitory (b) synapses by the colocalization of pre- and postsynaptic markers. Analyses were focused within

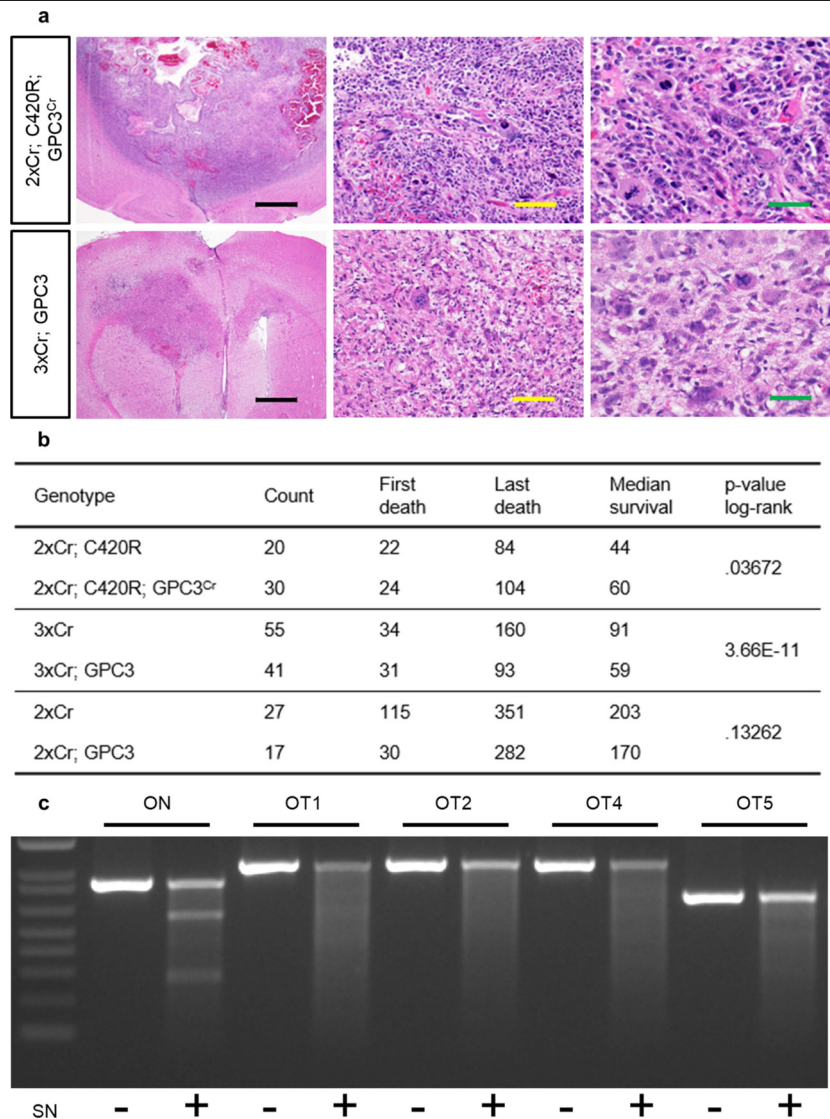
200 µm of the tumour margin (dotted line), as marked by GFP (pseudo-coloured in blue). Higher magnification images from the red and blue boxes are displayed. White scale bars, 200 µm; blue scale bars, 50 µm; yellow scale bars, 12.5 µm. Experiments were independently repeated 15 times.





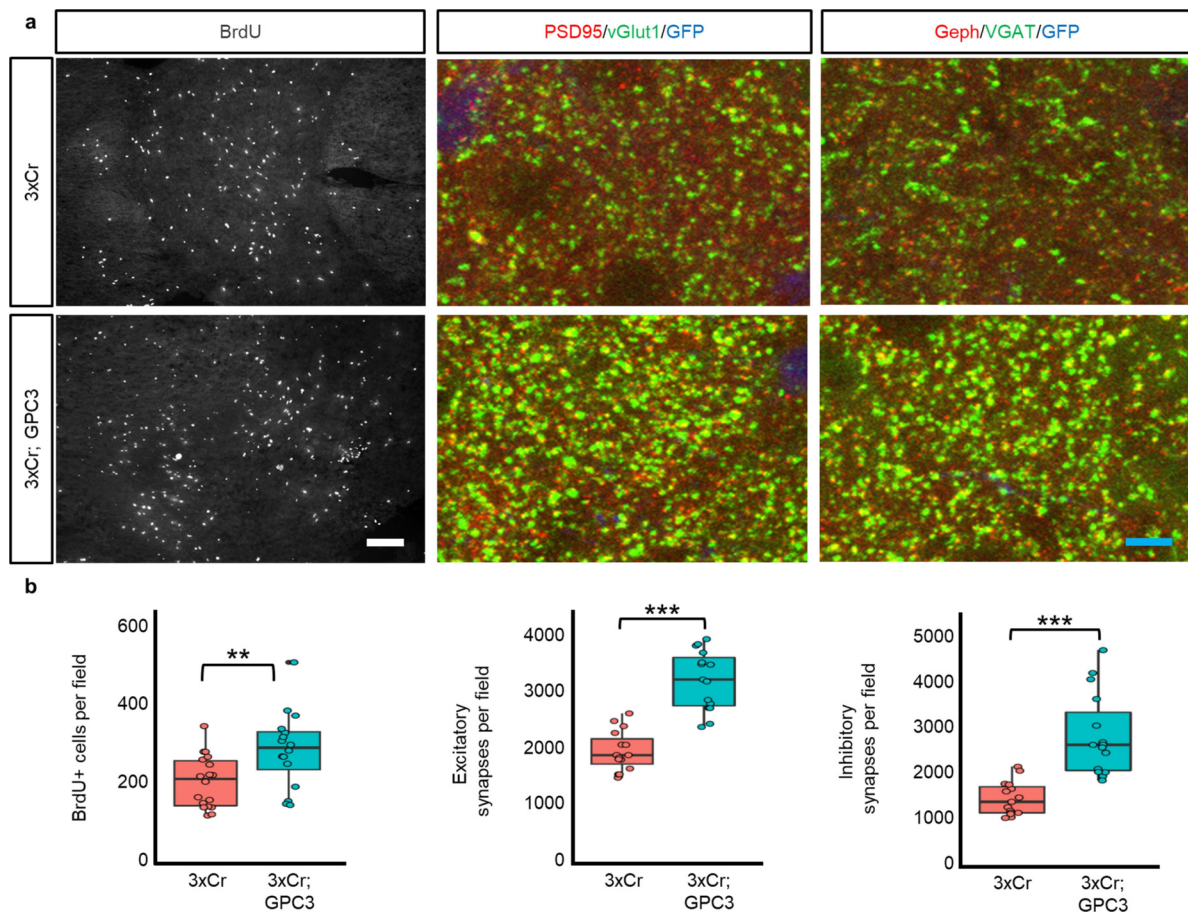
**Extended Data Fig. 6 | Expression of GPC family members across mouse and human glioblastoma models.** **a**, RNA-seq analysis for astrocyte-secreted factors that promote synaptogenesis. Each column represents an average of biological replicates.  $n = 2$  (wild-type),  $n = 4$  (C420R), and  $n = 3$  (H1047R) mice.  $P$  values were determined by two-sided  $t$ -test on log-transformed expression values. **b**, RNA-seq data from the TCGA comparing the same set of genes across glioblastoma (GBM) and low-grade glioma (LGG). Asterisks denote significant

difference between glioblastoma and low-grade glioma, determined by two-sided  $t$ -test on log-transformed expression values. **c**, Immunofluorescence analyses of P30 variant tumour brains stained for GPC3 (red) and GFP (green), demonstrating GPC3 staining at the tumour core (left) and tumour margin (right).  $n = 4$  mice for each variant. White scale bars, 100  $\mu$ m. Experiments were independently repeated four times.



**Extended Data Fig. 7 | Context-specific requirement for GPC3 in glioma tumorigenesis. a**, H&E staining of 2xCr; C420R; GPC3<sup>Cr</sup> (top) and 3xCr; GPC3 (bottom) tumour brains histologically graded as high-grade glioma (either WHO grade III anaplastic astrocytoma or grade IV glioblastoma). Black scale bars, 1 mm; yellow scale bars, 100  $\mu$ m; green scale bars, 50  $\mu$ m. Representative images of each variant driven tumour are from  $n = 6$  brains. **b**, Survival statistics for in vivo modelling of GPC3 loss and gain in various tumour models. *P* values were determined by log-rank test. **c**, SURVEYOR assay analysis of genomic DNA

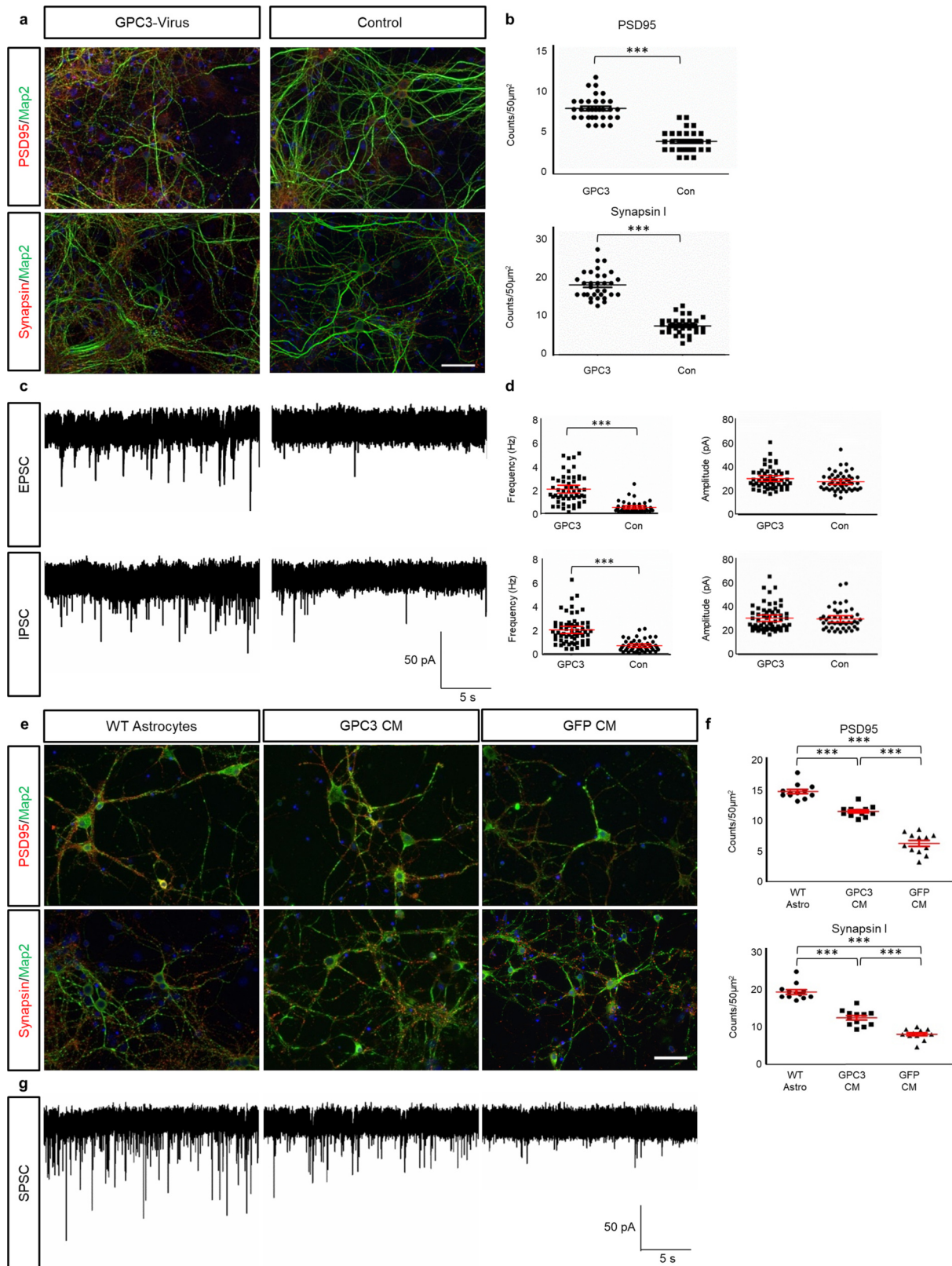
from 2xCr; C420R; GPC3<sup>Cr</sup> tumour tissue. Representative DNA gel electrophoresis of PCR products after SURVEYOR enzyme treatment. After the ladder (1-kb Plus, ThermoFisher, 10787018), which ranges from 100 to 1,500 bp, lanes from left to right contain on-target (ON) and top five off-target (OT1–OT5) sites with and without SURVEYOR nuclease (SN) treatment. OT3 rests in an AT-rich region and was not amplifiable. This experiment was independently repeated three times with similar results.



**Extended Data Fig. 8 | GPC3 promotes gliomagenesis and synaptic imbalance.** **a**, Immunofluorescence staining of 3xCr control and *Gpc3* overexpression tumours for BrdU, PSD95, VGLUT1 and VGAT. White scale bar,  $50 \mu\text{m}$ ; blue scale bar,  $5 \mu\text{m}$ . **b**, Quantification of BrdU immunofluorescence analysis (left) and excitatory (middle) and inhibitory (right) synapses.  $n = 4$

(BrdU) and  $n = 3$  (synapse) mice for each condition;  $n = 18$  (BrdU-3xCr),  $n = 16$  (BrdU-GPC3) and  $n = 15$  (synapse) technical repeats for each condition and synapse type. Field for BrdU incorporation =  $1,600 \mu\text{m}^2$ ; field for synapse analysis =  $34,000 \mu\text{m}^2$ . Box plots are as in Fig. 3b. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , one-way ANOVA.





**Extended Data Fig. 9** | See next page for caption.

## Extended Data Fig. 9 | GPC3 promotes synaptogenesis.

**a**, Immunofluorescence staining of co-cultures of astrocytes and neurons for PSD95, synapsin1 and MAP2 (Fig. 3b), with astrocytes overexpressing GPC3 via virus. Non-infected astrocytes were used as a control. White scale bar, 50  $\mu$ m.

**b**, Quantification of PSD95 (top) and synapsin1 (bottom) staining.  $n = 32$  technical repeats for each condition. Data are mean and s.e.m. \*\*\* $P < 0.001$ , one-tailed independent  $t$ -test; Tukey's test was used to compare individual mean values. **c**, Representative traces of from whole-cell recording of neurons co-cultured on astrocytes virally overexpressing GFP (control) or GPC3, with associated scale bar. **d**, Quantification of EPSC and IPSC amplitude and frequency from co-cultures.  $n = 54$  (GPC3, EPSC);  $n = 48$  (control, EPSC);  $n = 60$  (GPC3, IPSC);  $n = 46$  (control, IPSC). Data are mean and s.e.m. \*\*\* $P < 0.001$ ,

one-tailed independent  $t$ -test; Tukey's test was used to compare individual mean values. **e**, Immunofluorescence staining of neuron cultures for PSD95, synapsin1 and MAP2. Wild-type astrocyte–neuron co-cultures served as a positive control. Cortical neuron cultures were grown in GPC3 condition medium (CM) or GFP control conditioned medium. White scale bar, 50  $\mu$ m.

**f**, Quantification of PSD95 and synapsin1 staining.  $n = 12$  technical replicates for each condition. Data are mean and s.e.m. \*\*\* $P < 0.001$ , one-tailed independent  $t$ -test; Tukey's test was used to compare individual mean values. **g**, Representative traces of spontaneous postsynaptic current (SPSC) analysis of neurons co-cultured with astrocytes, GPC3 conditioned medium or GFP control medium, with associated scale bar.

Extended Data Table 1 | PIK3CA variant modelling survival statistics

Genotype	Count			Survival Range (days)			p-values (log rank test)								
	Total	♂	♀	First death	Last death	Median	2xCr; H1047R	2xCr; M1043I	2xCr; H701P	2xCr; V638A	2xCr; E545K	2xCr; C420R	2xCr; R88Q	2xCr; PIK3CA	2xCr; Cherry
3xCr	42	23	19	53	105	84	0.000129	0.137233	7.49E-11	1.73E-10	0.000733	1.61E-15	0.000605	6.65E-13	2.28E-13
2xCr; Cherry	28	12	16	115	351	203	4.51E-13	1.54E-12	0.000378	0.606653	2.56E-13	9.73E-15	6.69E-11	0.000469	
2xCr; PIK3CA	29	16	13	50	252	150	3.62E-13	6.34E-10	4.41E-09	0.00151	3.10E-13	1.02E-15	1.61E-06		
2xCr; R88Q	21	11	10	39	155	103	2.54E-05	0.067048	1.43E-10	2.36E-08	9.78E-05	2.71E-08			
2xCr; C420R	20	12	8	22	84	44	0.008385	8.79E-08	1.30E-11	1.60E-11	0.001131				
2xCr; E545K	32	17	15	37	119	63	0.398171	0.001046	9.81E-11	2.49E-10					
2xCr; V638A	20	11	9	99	>365*	203	3.90E-10	4.38E-09	0.000867						
2xCr; H701P	20	12	8	200	>365*	273	1.79E-10	1.17E-26							
2xCr; M1043I	19	12	7	51	134	80	0.000236								
2xCr; H1047R	30	13	17	32	115	59									

Statistics for survival study of PIK3CA variants in 2xCr background. 3xCr model served as a positive control. All studies were terminated at 1 year of age. P values were determined by log-rank test.

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

## Software and code

Policy information about [availability of computer code](#)

- |                 |  |
|-----------------|--|
| Data collection | Sequencing alignment and transcript abundance estimations were performed using HISAT (version 2.1.0) and Cufflinks (version 2.2.1.2). Combat software (PMID: 16632515) was used to alleviate any potential batch effect due to significant time between sequencing runs. |
| Data analysis   | Expression heat maps were generated using JavaTreeView (version 1.1.6r4) (PMID: 15180930).   |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in this study are available through the NCBI Gene Expression Omnibus (GEO) Repository with a series accession number GSE123519. Sample accession number range from GSM3506046 to GSM3506071, 26 samples total. Figure 2, extended data figures 5, supplemental figure 4, and supplemental table 3 present associated data.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No calculations were used to determine sample size for any of our experiments. In general, we aimed for a minimal sample size of 3 for significance. However, each experiment may have had its own minimal sample size for the varying reasons (abundant substrate, pragmatic rationale, based on other published work). Our reasons for samples size for each experiment are as follows: for our barcode sequencing screen, we initially analyzed 5 animals in technical triplicates. Across biological replicates, this generated strong reproductions thus we did not extend our analyses into more tissue. For our survival studies, we aimed for approximately 20-25 animals as each animal would then represent 4-5 percentage points of resolution. If more animals were generated for a particular genotype, they were included. Histological analysis was done on at least 6 tumors per variant. Sequencing of tumor tissues was done on at least 3 animals with the exception of our nontumor control which we only did in duplicate. Given how well each variant aligned within its own cohort compared to other variants, we felt confident more samples were not required. Both MRI and immunohistofluorescence (IHF) analysis were performed on 4 animals per condition. In vitro co-culture and electrophysiology were across biological triplicates. For these aforementioned experiments (MRI, IHF, coculture), given the minimal N of 3 and the statistically significant trends we observed, we did not see a need for increased sample size. EEG analysis was one 4 animals per genotype at 5 or 10 day intervals. Other studies, even with smaller variations to test, do not demonstrate the rigor in temporal profiling that we executed. Given this and the statistically significant trends we observed, we did not see the need to increase our sample size.
Data exclusions	The only samples that were excluded from this study were mouse that did not have any GFP reporter activity. Our method utilizes a GFP report as a control for proper technique. The absence of GFP would be indicative of poor electroporation technique. Only these samples were excluded.
Replication	Mouse samples were obtained across multiple litters. Within a single genotype, electroporation were done in at least 2 pregnant mothers on different dates. These cohorts were analyzed separately and compared. Only when these two cohorts were not significantly different, did we merge them together. Through the course of our study, replicate cohorts were never significantly different ( $p > .05$ ). All attempts at replication were successful.
Randomization	While we reason randomization was not necessary, we still exercise measures for randomization. For RNA-Sequencing, samples were taken from animals that expired across the whole range of their survival. The 3 animals were either among the first, middle, or last third to die. We utilized a similar approach in the samples used for our H&E stained histological analysis.
Blinding	While we reason that blinding was not necessary, we still exercised measures for blinding. This project also required much collaboration. Whenever samples were passed from one investigator to another, sample were only given a de-identified label. The samples were only properly labeled after analysis was complete.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input checked="" type="checkbox"/> MRI-based neuroimaging

## Antibodies

### Antibodies used

Antibodies are listed in the following format: host species anti-antigen (clone, dilution; source, product number; lot number). Clone IDs are not reported if one was not assigned.

For IHC and ICC:



rat anti-BrdU (BU1/75 (ICR1), 1:200; ABCAM, #ab6326; GR3269246-1), mouse anti-Gephyrin (3B11, 1:500; Synaptic Systems, 147011; 1-64), rabbit anti-GFP (1:1000; ThermoFisher, #A-11122; 1753594), goat anti-GPC3 (W-18, 1:100; Santa Cruz Biotechnology, #SC-10455; B2912), chicken anti-MAP2 (1:1,000; EnCor Biotech. Inc., #CPCA-MAP2; 7225-9), rabbit anti-synapsin-I (1:1,000; Millipore, #AB1543; 2363548), mouse anti-PSD95 (for ICC) (K28/43, 1:1,000; UC Davis/NIH NeuroMab Facility, #75-028; 455.7.JK.22), rabbit anti-human HLA A (EP1395Y, 1:100, ABCAM, ab52922; GR258732-22), mouse anti-PSD95 (for IHC) (7E3-1B8, 1:500; ThermoFisher, MA1-046; LJ147875), guinea pig anti-vGAT (1:500; Synaptic Systems, 131004; 2-41), and guinea pig anti-vGlut1 (1:2000; Millipore, AB5905; 3193844)

For RPPA, lot numbers were not available. These were used through the MD Anderson Functional Proteomics RPPA Core. Details stated in methods: rabbit anti-4E-BP1 phospho-S65 (1:250; Cell Signaling Technology, #9456), rabbit anti-Akt phospho-S473 (1:150; Cell Signaling Technology, #9271), rabbit anti-Akt phospho-T308 (1:250; Cell Signaling Technology, #2965), rabbit anti-GSK3 $\alpha/\beta$  phospho-S21/S9 (1:200; Cell Signaling Technology, #9331), rabbit anti-mTOR phospho-S2448 (1:50; Cell Signaling Technology, #2971), rabbit anti-p70 S6 Kinase phospho-T389 (1:50; Cell Signaling Technology, #9205), rabbit anti-PRAS40 phospho-T246 (1:500; ThermoFisher, #441100G), rabbit anti-Rictor phospho-T1135 (1:200; Cell Signaling Technology, #3806), rabbit anti-S6 phospho-S235/S236 (1:2500; Cell Signaling Technology, #2211), rabbit anti-S6 phospho-S240/S244 (1:1000; Cell Signaling Technology, #2215); rabbit anti-Tuberin/TSC2 phospho-T1462 (1:38; Cell Signaling Technology, #3617)

## Validation

All immunohistofluorescence and immunocytofluorescence staining were accompanied with secondary antibody only controls (not included in manuscript). Presented images are only those where accompanying secondary only controls showed no signal. Validation for these antibodies are as follows: rat anti-BrdU (PMID: 16670699), rabbit anti-GFP (currently over 266 published manuscripts listed: <https://www.thermofisher.com/antibody/product/GFP-Antibody-Polyclonal/A-11122>), goat anti-GPC3 (PMID: 21112773, 16734618, 15475451), chicken anti-MAP2 (PMID: 15642108, 2469170, 6120944, 16832065, 18988710, 22623668, 16817858, 16487970), rabbit anti-synapsin-I (PMID: 26060345, 26184109, 24633176, 24633867, 24638034, 25403753, 24719092, 25471559, 24184637, 24945922), mouse anti-PSD95 (currently over 54 published manuscripts listed: <https://www.labome.com/review/gene/human/PSD-95-antibody.html>), and rabbit anti-human HLA A (currently lists 37 published manuscripts listed: <https://www.abcam.com/hla-a-antibody-ep1395y-ab52922.html>). The mouse anti-Gephyrin, mouse anti-PSD95, guinea pig anti-vGAT, and guinea pig anti-vGlut1 used for immunohistological analysis of synapses were all according to PMID: 27615741. Antibodies used for RPPA were validated through the Functional Proteomics RPPA Core adhering to the following 3 criterion: A single or dominant band on western blotting is required and dynamic range and specificity is determined using: 1) peptides, phosphopeptides, growth factors, inhibitors, RNAi, cells with wide levels of expression including 330 cell lines under multiple conditions on a single array; 2) Pearson correlation coefficient between RPPA and western blotting of greater than 0.7 is required; and 3) reproducible results intra-slide or interslide are critical.

## Eukaryotic cell lines

### Policy information about cell lines

Cell line source(s)	COS7 - ATCC, CRL-1651; primary human GBM cell line
Authentication	COS7 cells was not authenticated; primary human GBM cell lines were received from Dr Frederick Lang, MD Anderson, Houston, TX
Mycoplasma contamination	Cells were not tested for mycoplasma contamination
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No commonly misidentified cell lines were used

## Animals and other organisms

### Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mus musculus, outbred ICR CD-1 IGS, both male and female mice were used in this study. These mice were all mice were less than 1 year old. Additionally, SCID-ICR male mice were used for xenograft studies. These mice were 6 weeks old at tumor transplantation. Xenografts were left to grow up to 6 weeks post transplants.
Wild animals	No wild animals were used
Field-collected samples	No samples were collected from the field
Ethics oversight	All procedures were approved by the Institutional Animal Care and Use Committee at Baylor College of Medicine and conform to the US Public Health Service Policy on Human Care and Use of Laboratory Animals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

### Policy information about studies involving human research participants

Population characteristics	We do not have this information. All human derived samples were received from Dr Frederick Lang, MD Anderson (PMID: 29016843)
----------------------------	---

## Recruitment

We do not have this information. All human derived samples were received from Dr Frederick Lang, MD Anderson (PMID: 29016843)

## Ethics oversight

Samples were collected through the auspices of Dr Frederick Lang's IRB protocol (LAB04-001). We only recieved samples post-de-identification. We were told these samples were published (PMID: 29016843). We were not given any information towards the identify of the source patient. Additionally, we have not conducted any experiments whereby we could ascertain such information.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Magnetic resonance imaging

### Experimental design

## Design type

Weekly longitudinal structural imaging

## Design specifications

T2 weighted low resolution and high resolution anatomical scans were acquired one time per week for four weeks unless the animal perished prior to the four week time point.

## Behavioral performance measures

N/A

### Acquisition

## Imaging type(s)

Structural Imaging

## Field strength

9.4T

## Sequence &amp; imaging parameters

For in vivo brain imaging of each mouse, three low resolution scans (axial, sagittal and coronal slice orientations) and one high resolution scan (axial orientation) were acquired in succession in a single imaging session. T2-weighted images were obtained using a fast spin echo pulse sequence with excitation and refocusing flip angles of 90 and 180° respectively at a rare factor of 8. In each low-resolution brain scan (TR/TE = 2500/33 ms), 10-12 slices per brain volume were scanned in a field of view (FOV) of 4.0 × 4.0 cm<sup>2</sup>, with a matrix size of 256 × 256 pixels yielding a spatial resolution of 0.156 × 0.156 mm<sup>2</sup> with a slice thickness of 1.0 mm and an interslice distance of 1.5 mm, taking two averages per slice with a scan time per brain volume of 2 min 40 s. Each high-resolution brain scan (TR/TE = 2500/36 ms) was taken with 14-18 slices per brain volume, scanned in a field of view (FOV) of 3.5 × 3.5 cm<sup>2</sup>, with a matrix size of 384 × 384 pixels yielding a spatial resolution of 0.091 × 0.091 mm<sup>2</sup> with a slice thickness of 0.75 mm and interslice distance of 1.0 mm, taking 10 averages per slice with a scan time per brain volume of 8 min.

## Area of acquisition

Multiple slices were taken that covered the whole brain

## Diffusion MRI

☐ Used

☒ Not used

### Preprocessing

## Preprocessing software

None

## Normalization

None

## Normalization template

None

## Noise and artifact removal

None

## Volume censoring

Visual inspection -- no significant motion artifacts were detected during imaging.

### Statistical modeling & inference

## Model type and settings

We performed univariate analysis on the relative tumor area. Relative tumor area was calculated as the product of the size of the aberrant growth along the dorsal-ventral and medial-lateral axes. This analysis was performed at a fixed location along the rostral-caudal axis. Analysis was done across fixed time points: 4, 5, and 6 weeks of age. Comparisons were done across genotypes. The analysis was across 4 mice per cohort unless the mice died prior to the acquisition of a datapoint.

## Effect(s) tested

No behavior effects were tested. The only differences across cohorts were genetic differences. All the animals were treated the same prior to, during, and after the MRI recording session.

Specify type of analysis: ☐ Whole brain ☒ ROI-based ☐ Both

MicroDicom, a free DICOM viewer and software, was used to visualize and analyze MRI images. Suspected tumor regions were identified in the last scan, prior to mice being euthanized. These anatomical regions were backtracked across time. Suspected tumor regions were based against age match control animals which did not have any manipulation done to them, as well as differences in symmetry between the two hemispheres of the brain. Relative tumor area was calculated as the product of the size of the aberration along the dorsal-ventral and medial-lateral axes. The same anatomical sections along the rostral-caudal axis was used for analysis.

Anatomical location(s)

Statistic type for inference  
(See [Eklund et al. 2016](#))

N/A

Correction

No correction methods were used

## Models & analysis

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Involvement in the study  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Functional and/or effective connectivity     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Graph analysis                               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Multivariate modeling or predictive analysis |

# Constructing protein polyhedra via orthogonal chemical interactions

<https://doi.org/10.1038/s41586-019-1928-2>

Received: 9 May 2019

Accepted: 19 November 2019

Published online: 22 January 2020

Eyal Golub<sup>1</sup>, Rohit H. Subramanian<sup>1</sup>, Julian Esselborn<sup>1</sup>, Robert G. Alberstein<sup>1</sup>, Jake B. Bailey<sup>1</sup>, Jerika A. Chiong<sup>1</sup>, Xiaodong Yan<sup>2</sup>, Timothy Booth<sup>2</sup>, Timothy S. Baker<sup>2</sup> & F. Akif Tezcan<sup>1,3\*</sup>

Many proteins exist naturally as symmetrical homooligomers or homopolymers<sup>1</sup>. The emergent structural and functional properties of such protein assemblies have inspired extensive efforts in biomolecular design<sup>2–5</sup>. As synthesized by ribosomes, proteins are inherently asymmetric. Thus, they must acquire multiple surface patches that selectively associate to generate the different symmetry elements needed to form higher-order architectures<sup>1,6</sup>—a daunting task for protein design. Here we address this problem using an inorganic chemical approach, whereby multiple modes of protein–protein interactions and symmetry are simultaneously achieved by selective, ‘one-pot’ coordination of soft and hard metal ions. We show that a monomeric protein (protomer) appropriately modified with biologically inspired hydroxamate groups and zinc-binding motifs assembles through concurrent Fe<sup>3+</sup> and Zn<sup>2+</sup> coordination into discrete dodecameric and hexameric cages. Our cages closely resemble natural polyhedral protein architectures<sup>7,8</sup> and are, to our knowledge, unique among designed systems<sup>9–13</sup> in that they possess tightly packed shells devoid of large apertures. At the same time, they can assemble and disassemble in response to diverse stimuli, owing to their heterobimetallic construction on minimal interprotein-bonding footprints. With stoichiometries ranging from [2 Fe:9 Zn:6 protomers] to [8 Fe:21 Zn:12 protomers], these protein cages represent some of the compositionally most complex protein assemblies—or inorganic coordination complexes—obtained by design.

Cage-like architectures have featured prominently in supramolecular design, owing to their aesthetically appealing structures and isolated interiors, which enable them to encapsulate molecular cargo and to perform selective chemical transformations<sup>14–18</sup>. Inspired by naturally occurring polyhedral assemblies, protein engineers have combined principles of symmetry with the proper design and arrangement of non-covalent interfaces to build diverse supramolecular architectures<sup>9–13</sup>. However, some of the key structural features of natural protein cages have been difficult to emulate (Fig. 1a). First, each cage is invariably composed of asymmetric protomers, which possess multiple self-associative patches to simultaneously satisfy the symmetry requirements necessary to build polyhedral assemblies (that is, concurrent generation of at least  $C_2$  and  $C_3$  symmetries, in addition to  $C_4$  or  $C_5$  symmetries for octahedra or icosahedra)<sup>1,6</sup>. Second, these self-associative patches collectively occupy a large fraction of the surface area on each protomer, enabling the formation of tightly packed shells with small apertures to enable the influx and efflux of select species<sup>8</sup>. Third, although the inter-protomer interfaces in natural protein cages are extensive, to ensure stable and selective association, they are also often conformationally flexible and chemically tunable, allowing the cages to undergo cooperative motions or disassembly in response to external cues<sup>7,19</sup>.

Given the difficulty of designing multiple, selectively associative surfaces on a protomer, construction of artificial cages has relied

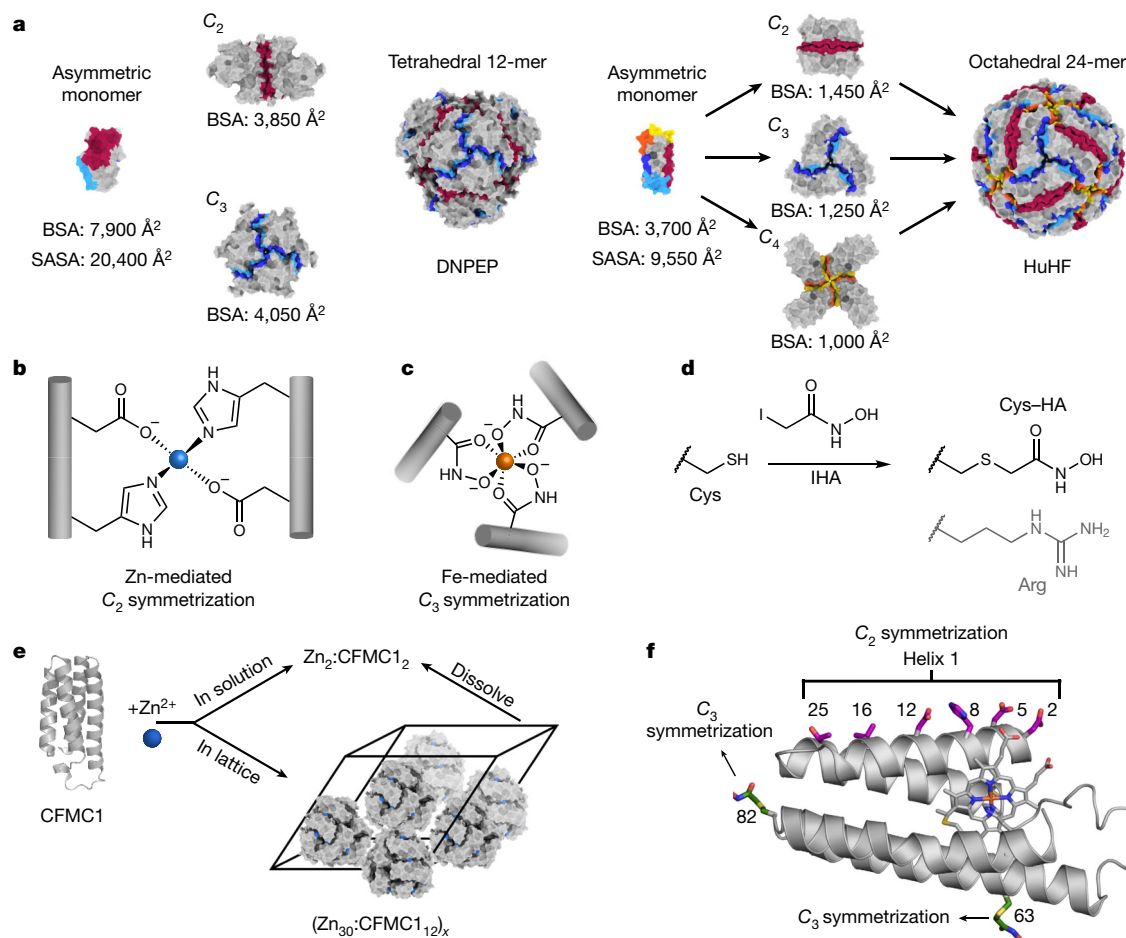
exclusively on using natively oligomeric proteins or designed peptides with  $C_{n \geq 2}$  symmetries as building blocks and the design of a single type of binary protein–protein interaction (PPI) through computation<sup>10</sup>, genetic fusion<sup>9,11</sup>, disulfide bond formation<sup>12</sup> or metal coordination<sup>11,13</sup>. Although these strategies can yield polyhedral symmetries, the resulting architectures are highly porous, do not display externally controllable assembly or disassembly (with two exceptions)<sup>11,13</sup> and cannot be easily modified to adopt alternative structures (that is, they are not modular or flexible). Inspired by previous work on bimetallic supramolecular coordination cages<sup>20,21</sup>, we investigated whether these design problems could be addressed using an inorganic chemical approach, wherein a protomer is equipped with chemically orthogonal coordination motifs to self-assemble into polyhedral architectures.

## Design of bimetallic protein cages

Previously, we have taken advantage of the simultaneous strength, lability and directionality of metal coordination bonds (particularly those formed by late first-row, low-valent transition metal ions) to effect the self-assembly of discrete protein complexes<sup>5</sup> and extended one-, two- and three-dimensional arrays<sup>22,23</sup>. Typically, selective nucleation sites for metal-mediated PPIs are formed by pairs of metal-binding amino acids (mainly His, Asp and Glu residues) (Fig. 1b) or non-native bidentate

<sup>1</sup>Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA. <sup>2</sup>Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA.

<sup>3</sup>Materials Science and Engineering, University of California, San Diego, La Jolla, CA, USA. \*e-mail: [tezcan@ucsd.edu](mailto:tezcan@ucsd.edu)



**Fig. 1 | Design of protein cages.** **a**, Representative examples of natural protein cages (DNPEP, aspartyl aminopeptidase; HuHF, human heavy-chain ferritin) and their assembly from asymmetric protomers. Per-protomer solvent-accessible surface areas (SASA) and buried surface areas (BSA) are indicated. Associative surfaces on the protomers are coloured red for homologous interactions and orange/yellow or blue/cyan for heterologous interactions. **b–f**, Design of artificial protein cages by metal coordination. **b**,  $C_2$ -symmetric

protein dimerization induced by tetrahedral  $Zn^{2+}$  coordination of native amino acid side chains. **c**,  $C_3$ -symmetric protein trimerization induced by octahedral  $Fe^{3+}$ -tris-hydroxamate coordination. **d**, Scheme showing modification of native Cys side chains with IHA to yield Cys-HA, which is isosteric with arginine (light grey). **e**, Zn-mediated solution dimerization and crystallization of CFMC1. **f**, Structural overview of the cytochrome  $cb_{562}$  scaffold. Salient structural elements are shown as sticks.

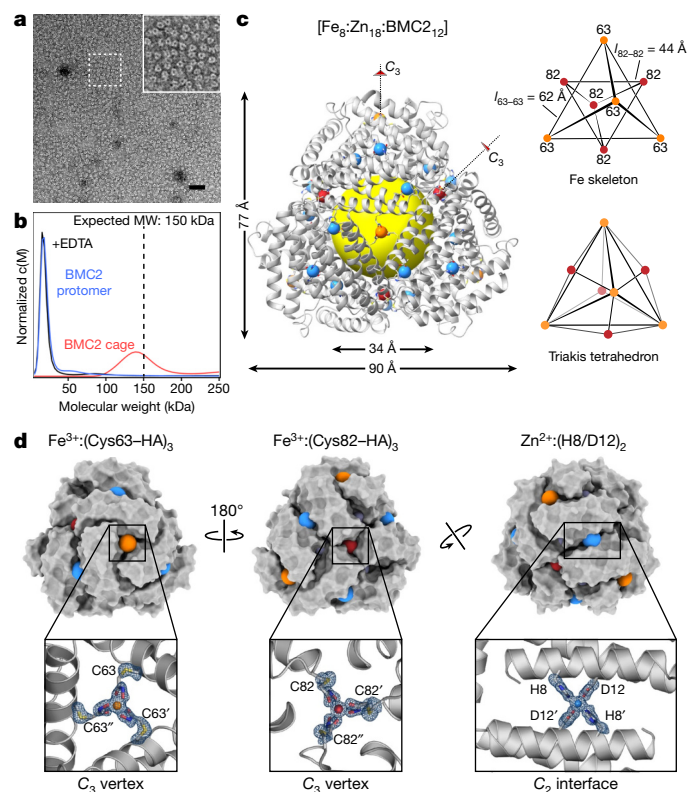
functionalities (for example, 2,2'-bipyridine, 1,10-phenanthroline and 8-hydroxyquinoline)<sup>5,24</sup>. However, all of these natural or synthetic coordination motifs can be considered as soft (or intermediate-soft) according to the Hard-Soft Acid-Base (HSAB) classification<sup>25</sup> and have considerable overlap in terms of their coordination preferences for soft, low-valent transition metal ions. Owing to this lack of chemical discrimination, it has not been possible to design a heterometallic protein complex for which the self-assembly is selectively guided by multiple metal ions that mediate different PPIs.

To achieve this goal, we turned to a bidentate chelating motif, hydroxamate (HA, the conjugate base of hydroxamic acid), a common functional group found in bacterial siderophores to enable exceptionally stable coordination of  $Fe^{3+}$  ions<sup>26,27</sup>. HA groups preferentially form octahedral  $Fe^{3+}$  complexes with an inherent  $C_3$  symmetry that we sought to impose on protein oligomerization (Fig. 1c). Notably, the formation constants of  $Fe^{3+}$ :(HA)<sub>3</sub> complexes ( $>10^{28} M^{-3}$ ) are vastly higher than those of other metal-HA complexes, such that they can be considered as orthogonal to the aforementioned soft metal-ligand combinations<sup>26,27</sup>. For protein derivatization, we synthesized a small reagent, iodo-hydroxamic acid (IHA), which selectively reacts with Cys residues (Fig. 1d, Extended Data Fig. 1). The resulting Cys-HA side chain is isosteric with that of arginine and devoid of bulky aromatic moieties, furnishing a pseudo-natural amino acid functionality with the ability

to chelate hard metal ions and induce  $C_3$  symmetric oligomerization on a single-residue footprint.

As a model system, we used cytochrome  $cb_{562}$ , a monomeric four-helix-bundle protein that has proved to be a versatile building block for metal-directed protein self-assembly<sup>5</sup>. A variant of cyt  $cb_{562}$  (CFMC1), which was designed and observed to form Zn-mediated dimers in solution, crystallizes into rhombohedral lattices in which the protomers arrange into dodecameric, cage-like units via Zn-mediated crystal packing interactions<sup>28</sup> (Fig. 1e). Whereas Zn-mediated interactions were not sufficiently strong to maintain the tetrahedral dodecamers upon crystal dissolution, we envisioned that these lattice units could serve as a structural model to engineer the protomers such that they would form self-standing cages. Looking first to stabilize the  $C_2$  symmetric interfaces, we incorporated a bidentate His8-Asp12 motif to mediate the antiparallel association of two protomers along their helices 1 via tetrahedral  $Zn^{2+}$  coordination (Fig. 1f). Given that  $C_3$  symmetric interfaces are small and heterologous (that is, they involve two different patches on each protomer; Extended Data Fig. 2a), they were unsuitable for stabilization by noncovalent interactions. Therefore, we focused on the central pores in each  $C_3$  symmetric substructure and identified positions 63 and 82 as suitable locations for installing Cys-HA functionalities, which would stabilize trimeric substructures by forming  $Fe^{3+}$ :(HA)<sub>3</sub> centres (Fig. 1f). Thus, we prepared two CFMC1 variants





**Fig. 2 | Characterization of BMC2 cages.** **a**, ns-TEM of BMC2 cages obtained by the dissolution of three-dimensional crystals. Inset, close-up of the boxed region. Scale bar, 50 nm. **b**, AUC characterization of BMC2 protomers and BMC2 cages after crystal dissolution and after subsequent treatment with EDTA. **c**, Crystal structure of the BMC2 cage. Fe and Zn ions are represented as orange/red and blue spheres, respectively. The central cavity is highlighted by a yellow sphere. Two types of  $\text{C}_3$  vertices formed by  $\text{Fe}:(\text{Cys63-HA})_3$  and  $\text{Fe}:(\text{Cys82-HA})_3$  coordination motifs form two superimposed tetrahedra to generate a triakis tetrahedron. **d**, Surface representations of the BMC2 cage, with metal ions shown as coloured spheres. Insets show atomic details of each metal coordination site, with the  $mF_o - DF_c$  electron density omit map (blue mesh) contoured at  $3\sigma$ .

designated bimetallic cage 1 and bimetallic cage 2 (BMC1 and BMC2; Extended Data Fig. 2b). Both BMC1 and BMC2 bear the His8–Asp12 motif on helix 1 and Cys63–HA along with the native peripheral Zn coordination sites (Ala1<sub>N-term</sub>, Asp39 and His77) of the parent CFMC1 structure. BMC2 additionally contains Cys82–HA (Extended Data Figs. 1, 2b).

Crystals of BMC1 and BMC2 were obtained in the presence of near equimolar  $\text{ZnCl}_2$  and  $\text{FeSO}_4$ . These crystals were isomorphous ( $R32$  space group;  $a = b = 126 \pm 1$  Å,  $c = 167 \pm 1$  Å) with those of CFMC1<sup>28</sup>, indicating that they possessed the same underlying lattice structure composed of dodecameric units (Extended Data Table 1, Extended Data Fig. 2). Crystals were dissolved in a solution lacking the precipitating agent (PEG-400) and then analysed by negative-stain transmission electron microscopy (ns-TEM; Fig. 2a, Extended Data Fig. 3). The images revealed uniform particles with a diameter of  $8.4 \pm 0.8$  nm in the case of BMC2 but not BMC1, implying that two HA coordination motifs are necessary for cage stability. Analysis of the same BMC2 solution by analytical ultracentrifugation (AUC) indicated a predominant species with a molecular weight ( $\text{MW}_{\text{obs}}$ ) of about 140 kDa (Fig. 2b), approximating the calculated value ( $\text{MW}_{\text{calc}}$ ) of 150 kDa for a dodecamer. BMC2 particles dissociated upon treatment with ethylenediamine tetraacetic acid (EDTA), confirming their metal-dependent self-assembly (Fig. 2b, Extended Data Fig. 3).

We determined the crystal structure of the BMC2 cage at 1.4 Å resolution (Extended Data Table 1), revealing a compact structure with

the shape of a truncated tetrahedron, outer dimensions of  $80 \times 90$  Å and a cavity volume of  $32,700$  Å<sup>3</sup> (Fig. 2c, Extended Data Fig. 4). Like natural protein cages, the shell is tightly packed and the largest opening measures less than 4 Å across. Nearly 30% of the surface area of each protomer ( $1,700$  Å<sup>2</sup> out of  $6,500$  Å<sup>2</sup>) is buried in interfaces despite a design footprint of only four amino acids (His8, Asp12, Cys63 and Cys82).

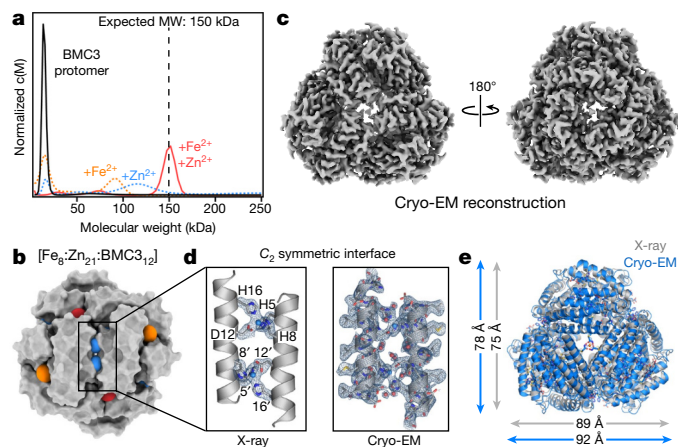
The full complement of metal ions, comprising eight Fe ions (four each in the  $\text{C}_3$  symmetric pores) and eighteen Zn ions (six in  $\text{C}_2$  interfaces and twelve in peripheral sites) are clearly resolved (Fig. 2c). Anomalous X-ray diffraction data collected at and below Fe and Zn K-edges indicate that the designed Fe- and Zn-coordination sites exclusively bind to their cognate ions with no evidence of crosstalk (Extended Data Fig. 5, Supplementary Tables 3–6), which establishes that the metal-dependent self-assembly of BMC2 cages occurs with absolute chemical selectivity. The Fe centres form the eight  $\text{C}_3$  vertices of a triakis tetrahedron, a Catalan solid with twelve equivalent faces (Fig. 2c). It can be viewed as the superposition of two tetrahedra: four Fe centres that are coordinated by Cys63–HA motifs generate the larger of these two tetrahedra (with an edge length ( $l_{\text{edge}}$ ) of 62 Å), and four Fe centres coordinated by Cys82–HA motifs produce the smaller one ( $l_{\text{edge}} = 44$  Å) (Fig. 2c). The BMC1 structure, in comparison, has a regular tetrahedral arrangement of four Fe centres as it lacks the Cys82–HA group (Extended Data Table 1, Extended Data Fig. 2).

As designed, the edges of the BMC2 tetrahedra are formed by six Zn ions located centrally in  $\text{C}_2$  interfaces (Fig. 2d). The  $\text{Fe}^{3+}:(\text{Cys63-HA})_3$  and  $\text{Fe}^{3+}:(\text{Cys82-HA})_3$  motifs display near-ideal octahedral geometries (Fig. 2d), with the former in  $\Lambda$  (left handed) and the latter in  $\Delta$  (right handed) configuration (Extended Data Fig. 5). Notably, the  $\text{Fe}^{3+}:(\text{Cys82-HA})_3$  centre also adopts an alternative conformation (20% abundance) owing to the flexibility of the Cys–HA side chain (Extended Data Fig. 5k). All Fe–O bond distances are in the range of 1.95–2.1 Å, which are typical of  $\text{Fe}^{3+}:(\text{HA})_3$  complexes<sup>29</sup>. Given that a  $\text{Fe}^{2+}$  precursor was used to initiate self-assembly, and as  $\text{Fe}:(\text{HA})_3$  centres have low reduction potentials ( $E_{\text{red}} < -400$  mV)<sup>26</sup>, this observation suggests that the protein self-assembly involves the initial formation of  $\text{Fe}^{2+}:(\text{HA})_3$  centres, followed by the thermodynamically favoured oxidation of these species into  $\text{Fe}^{3+}$  either by the  $\text{Fe}^{3+}$ –haem centres embedded in each protomer (see Methods) or directly by ambient  $\text{O}_2$ .

## Reversible assembly of dodecameric cages

Despite the stability of isolated BMC2 cages, their formation required an initial crystallization step. We reasoned that slow crystal nucleation or growth kinetics and the high attendant protein and metal concentrations probably increased the fidelity and yield of the complex self-assembly process to produce a discrete supermolecule consisting of 12 protomers and 26 metal ions of two different kinds. Reasoning that strengthened Zn-mediated interactions across the  $\text{C}_2$  interface could increase the efficiency of cage self-assembly in solution, we generated two second-generation variants based on BMC2: BMC3 and BMC4 (Extended Data Figs. 1, 2b). In BMC3, the helix 1 surface was engineered to form two Zn-coordination sites (composed of His5, His8, Asp12 and His16) across the  $\text{C}_2$  interface, whereas in BMC4 three potential Zn-coordination sites were engineered (one central site composed of two His8–Asp12 pairs as in BMC3 and two peripheral sites composed of Glu2, Glu5, His16 and Glu25). In BMC4, we also removed the Cys63–HA group with the purpose of eliminating any potential undesired assembly products that involve heteromeric  $\text{Fe}^{3+}$  coordination by Cys63–HA and Cys82–HA.

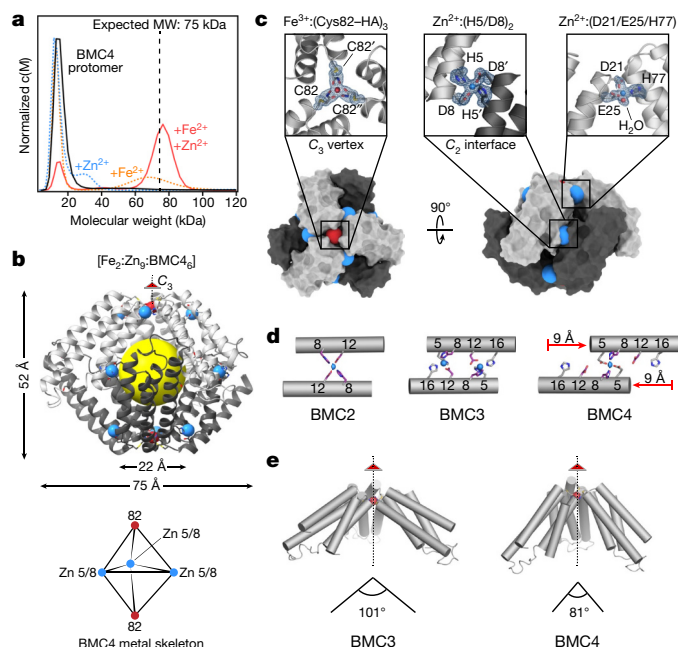
BMC3 indeed formed dodecameric cages in solution with high yields (>80%) as determined by ns-TEM and AUC measurements (Fig. 3a, Extended Data Fig. 3). The 1.85 Å-resolution crystal structure confirmed the eight Fe centres in the vertices and the two Zn coordination sites in each  $\text{C}_2$  interface (twelve in total) as well as nine of the twelve possible



**Fig. 3 | Characterization of BMC3 cages.** **a**, AUC characterization of BMC3 self-assembly. **b**, Surface representation of the BMC3 cage (as derived from the crystal structure), oriented to show the incorporation of two Zn ions at the  $C_2$  symmetric interface. **c**, The 2.6 Å-resolution cryo-EM electron density map for the BMC3 cage. **d**, Atomic details of both Zn-binding sites of the two-fold interface overlaid with the electron density  $mF_o - DF_c$  omit map from the crystal structure (left) and as observed for the cryo-EM structure (right). Additional side chains and waters are shown for the cryo-EM structure to emphasize the structural robustness of the interface. **e**, Overlay of the BMC3 X-ray and cryo-EM structures to highlight the isotropic expansion of the cage in the absence of crystallographic packing interactions.

peripheral Zn sites, which complete a [8 Fe:21 Zn:12 protomers] architecture (Fig. 3b). Notably, the self-assembly of BMC3 cages in solution was dependent on the presence of both Fe and Zn ions. The absence of either metal ion or the addition of various other first-row transition metal ions instead of Fe led to smaller oligomeric forms of BMC3 or non-specific assemblies (Fig. 3a, Extended Data Fig. 6a). BMC3 cages also formed with a  $Fe^{3+}$  precursor, Fe(acetylacetonate)<sub>3</sub> (Extended Data Fig. 6b). Consistent with self-assembly under thermodynamic control, the formation of dodecameric cages was independent of the order of addition of Fe or Zn ions. We determined the single-particle cryo-electron microscopy (cryo-EM) structure of isolated BMC3 cages at a resolution of 2.6 Å (Fig. 3c, Extended Data Table 2). A major portion of the assembly could be resolved at 2.0 Å or less (Extended Data Fig. 7). At this resolution, nearly all side chains, Zn coordination sites and some ordered water molecules are clearly distinguished (Fig. 3d). Consistent with the crystallographically observed flexibility of  $Fe^{3+}:(Cys-HA)_3$  coordination sites, electron densities in the  $C_3$  vertices are diffuse and some side chains that display high temperature factors in the crystal structure are found in alternative conformations in the cryo-EM structure (Extended Data Fig. 7). These observations confirm that the solution architecture of the BMC3 cage closely reflects the solid-state structure. Probably owing to lattice packing, the latter is isotropically compressed by around 2–3 Å compared to the former (Fig. 3e), which can be accommodated by slight changes in interfacial metal coordination.

Next, we examined the assembly and disassembly behaviour of BMC3 cages in response to different stimuli. BMC3 cages readily disassemble upon treatment with EDTA (Extended Data Fig. 6c). They were stable at 50 °C but dissociated upon incubation at 70 °C (Extended Data Fig. 6d). A key feature of siderophores is that their cellular release of Fe is promoted by the destabilization and labilization of their  $Fe^{3+}:(HA)_3$  centres through reduction to the  $Fe^{2+}$  form in the cytosol<sup>27</sup>. Along these lines, the treatment of BMC3 cages with a strong reductant (dithionite;  $E_{red} < -500$  mV)<sup>30</sup> led to their disappearance and the emergence of monomeric species (Extended Data Fig. 6e). By contrast, a weaker reductant (ascorbate;  $E_{red} > -100$  mV at pH 7)<sup>31</sup> with a reduction potential higher than that of  $Fe^{3+}:(HA)_3$  had considerably less effect (Extended Data Fig. 6e), suggesting that the disassembly of BMC3 cages occurs



**Fig. 4 | Characterization of BMC4 cages.** **a**, AUC characterization of BMC4 self-assembly. **b**, Crystal structure of the BMC4 cage. Fe and Zn ions are represented as red and blue spheres, respectively. The central cavity is highlighted by a yellow sphere. The structural skeleton formed by Fe and Zn ions is shown below the structure. **c**, Surface representations of the BMC4 cage, with metal ions shown as coloured spheres. Atomic details of each metal coordination site are shown in insets, with the  $mF_o - DF_c$  electron density omit maps (blue mesh) contoured at  $3\sigma$ . **d**, Comparison of the  $C_2$  symmetric protein interfaces in different BMC constructs. Residues 8 and 12, which are common to all constructs, are coloured purple. The slippage of the two-fold helix interface to accommodate the hexameric architecture of BMC4 is indicated with red arrows. **e**, Comparison of the apical angle formed by the  $Fe:(Cys82-HA)_3$ -mediated vertices in BMC3 and BMC4 cages.

through the reduction of the Fe centres. These observations establish BMC3 cages as a distinctive system among natural and artificial protein architectures the assembly and disassembly of which can be controlled through multiple stimuli: chemical, thermal or redox. BMC3 cages can passively encapsulate small fluorogenic molecules in either their lumen or inter-protomer interfaces, retain them for several days and release them upon treatment with EDTA (Extended Data Fig. 8).

## Formation of a hexameric cage

Unexpectedly, AUC measurements indicated that the other second-generation variant, BMC4, self-assembled as a hexamer upon Fe and Zn coordination, with yields exceeding 70% (Fig. 4a). The 1.50 Å resolution crystal structure of the BMC4 assembly revealed a  $D_3$  symmetric, cage-like architecture with a composition of [2 Fe:9 Zn:6 protomers], outer dimensions of 75 Å × 50 Å and a cavity volume of more than 7,800 Å<sup>3</sup> (Extended Data Fig. 4). The overall shape is a trigonal bipyramid (Fig. 4b), which is the smallest polyhedral architecture with a sizeable interior cavity that can be constructed from an asymmetric building block. The apical vertex of each pyramidal half is formed by a  $Fe^{3+}:(Cys82-HA)_3$  motif shared by three protomers (Fig. 4b). These  $C_3$  symmetric vertices are further reinforced by  $Zn^{2+}$  ions that link pairs of protomers through Asp21, Glu25 and His77 coordination (Fig. 4c). The pyramids are joined by three equatorial,  $C_2$  symmetric vertices mediated by Zn centres coordinated to Glu5 and His8 (Fig. 4b, c). A comparison to the BMC2 and BMC3 cages indicates that this unexpected coordination motif requires an approximately 9 Å slip of each protomer along the  $C_2$  symmetric interfaces (Fig. 4d). The shift markedly reduces

the  $C_2$  symmetric contact area between protomers, effectively transforming the edges in the tetrahedral BMC2 and BMC3 cages to vertices in the trigonal bipyramidal BMC4 cages. BMC4 cages exhibited similar thermal stability to BMC3 cages, with both species disassembling at below 70 °C. The thermal robustness of the BMC3 and BMC4 cages appear to be limited, at least in part, by the relative instability of the individual protomers (Extended Data Fig. 6d).

The large structural transformation is accompanied by a reduction in the apical angle formed at the  $\text{Fe}^{3+}:(\text{Cys82-HA})_3$ -mediated vertices from 101° in the BMC2 and BMC3 cages to 81° in the BMC4 cage (Fig. 4e). This observation highlights the conformational adaptability of the  $\text{Fe}^{3+}:(\text{Cys82-HA})_3$  coordination motif, enabling it to accommodate different polyhedral geometries. Such behaviour is reminiscent of the interfacial flexibility in some icosahedral virus capsids in which the same protomer can form hexamers on capsid faces and pentamers on capsid vertices<sup>7</sup>. It is worth noting that BMC4 contains all of the Zn-coordinating residues on helix 1 to form the  $C_2$  symmetric interfaces observed in the dodecameric BMC3 cage, indicating that the self-assembly process selects an alternative interfacial arrangement of lower free energy, enabled by the reversibility of metal coordination interactions. In terms of protein design, a caveat of interfacial flexibility is that it may lead to nonspecific or unintended self-assembly products, although it can also allow error correction during self-assembly and increase tolerance to design imperfections.

## Conclusions

The self-assembly and function of biomolecular systems are predicated upon their specificity, stability and adaptiveness, which, in turn, are enabled by extensive networks of non-covalent interactions. Here, we have shown that fundamental concepts in inorganic coordination chemistry can be applied to achieve all of these attributes in protein self-assembly and, specifically, to construct complex polyhedral protein architectures from a simple, asymmetric building block. Despite their minimal design footprints, these cage-like architectures are distinguished by their structural compactness and responsiveness—hallmarks of evolved systems such as viral capsids. Key to our construction strategy was the reimagining of a biological coordination motif, hydroxamic acid, within a new structural context: as a new amino acid side chain with the ability to chelate hard metal ions. This example expands the growing lexicon of post-translational modifications that broaden the chemical scope of proteins.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1928-2>.

1. Marsh, J. A. & Teichmann, S. A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* **84**, 551–575 (2015).

2. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl Acad. Sci. USA* **98**, 2217–2221 (2001).
3. Bai, Y., Luo, Q. & Liu, J. Protein self-assembly via supramolecular strategies. *Chem. Soc. Rev.* **45**, 2756–2767 (2016).
4. Hamley, I. W. Protein assemblies: nature-inspired and designed nanostructures. *Biomacromolecules* **20**, 1829–1848 (2019).
5. Churchfield, L. A. & Tezcan, F. A. Design and construction of functional supramolecular metalloprotein assemblies. *Acc. Chem. Res.* **52**, 345–355 (2019).
6. Yeates, T. O. Geometric principles for designing highly symmetric self-assembling protein nanomaterials. *Annu. Rev. Biophys.* **46**, 23–42 (2017).
7. Johnson, J. E. & Speir, J. A. Quasi-equivalent viruses: a paradigm for protein assemblies. *J. Mol. Biol.* **269**, 665–675 (1997).
8. Lawson, D. M. et al. Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. *Nature* **349**, 541–544 (1991).
9. Lai, Y.-T., Cascio, D. & Yeates, T. O. Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129 (2012).
10. Bale, J. B. et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016).
11. Cristie-David, A. S. & Marsh, E. N. G. Metal-dependent assembly of a protein nano-cage. *Protein Sci.* **28**, 1620–1629 (2019).
12. Fletcher, J. M. et al. Self-assembling cages from coiled-coil peptide modules. *Science* **340**, 595–599 (2013).
13. Malay, A. D. et al. An ultra-stable gold-coordinated protein cage displaying reversible assembly. *Nature* **569**, 438–442 (2019).
14. Pluth, M. D., Bergman, R. G. & Raymond, K. N. Acid catalysis in basic solution: a supramolecular host promotes orthoformate hydrolysis. *Science* **316**, 85–88 (2007).
15. Chakrabarty, R., Mukherjee, P. S. & Stang, P. J. Supramolecular coordination: self-assembly of finite two- and three-dimensional ensembles. *Chem. Rev.* **111**, 6810–6918 (2011).
16. Yoshizawa, M., Klosterman, J. K. & Fujita, M. Functional molecular flasks: new properties and reactions within discrete, self-assembled hosts. *Angew. Chem. Int. Ed.* **48**, 3418–3438 (2009).
17. Mal, P., Breiner, B., Rissanen, K. & Nitschke, J. R. White phosphorus is air-stable within a self-assembled tetrahedral capsule. *Science* **324**, 1697–1699 (2009).
18. Liu, Y., Hu, C., Comotti, A. & Ward, M. D. Supramolecular Archimedean cages assembled with 72 hydrogen bonds. *Science* **333**, 436–440 (2011).
19. Mateu, M. G. Assembly, stability and dynamics of virus capsids. *Arch. Biochem. Biophys.* **531**, 65–79 (2013).
20. Sun, X., Johnson, D. W., Caulder, D. L., Raymond, K. N. & Wong, E. H. Rational design and assembly of  $\text{M}_2\text{M}_3\text{L}_6$  supramolecular clusters with  $C_{3h}$  symmetry by exploiting incommensurate symmetry numbers. *J. Am. Chem. Soc.* **123**, 2752–2763 (2001).
21. Smulders, M. M. J., Jiménez, A. & Nitschke, J. R. Integrative self-sorting synthesis of a  $\text{Fe}_9\text{Pt}_6\text{L}_{24}$  cubic cage. *Angew. Chem. Int. Ed.* **51**, 6681–6685 (2012).
22. Brodin, J. D. et al. Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. *Nat. Chem.* **4**, 375–382 (2012).
23. Suzuki, Y. et al. Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals. *Nature* **533**, 369–373 (2016).
24. Radford, R. J., Nguyen, P. C. & Tezcan, F. A. Modular and versatile hybrid coordination motifs on  $\alpha$ -helical protein surfaces. *Inorg. Chem.* **49**, 7106–7115 (2010).
25. Pearson, R. G. Hard and soft acids and bases. *J. Am. Chem. Soc.* **85**, 3533–3539 (1963).
26. Wong, G. B., Kappel, M. J., Raymond, K. N., Matzkanke, B. & Winkelmann, G. Coordination chemistry of microbial iron transport compounds. 24. Characterization of coprogen and ferrirocen, two ferric hydroxamate siderophores. *J. Am. Chem. Soc.* **105**, 810–815 (1983).
27. Crumbliss, A. L. Iron bioavailability and the coordination chemistry of hydroxamic acids. *Coord. Chem. Rev.* **105**, 155–179 (1990).
28. Ni, T. W. & Tezcan, F. A. Structural characterization of a microperoxidase inside a metal-directed protein cage. *Angew. Chem. Int. Ed.* **49**, 7014–7018 (2010).
29. Failes, T. W. & Hambley, T. W. Crystal structures of tris(hydroxamate) complexes of iron(III). *Aust. J. Chem.* **53**, 879–881 (2000).
30. Mayhew, S. G. The redox potential of dithionite and  $\text{SO}_2$  from equilibrium reactions with flavodoxins, methyl viologen and hydrogen plus hydrogenase. *Eur. J. Biochem.* **85**, 535–547 (1978).
31. Borsook, H. & Keighley, G. Oxidation-reduction potential of ascorbic acid (vitamin C). *Proc. Natl Acad. Sci. USA* **19**, 875–878 (1933).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

# Retraction Note: Microglia-dependent synapse loss in type I interferon-mediated lupus

---

<https://doi.org/10.1038/s41586-020-1949-x>

---

Retraction to: *Nature* <https://doi.org/10.1038/nature22821>

---

Published online 14 June 2017

---

Allison R. Bialas, Jessy Presumey, Abhishek Das,  
Cees E. van der Poel, Peter H. Lapchak, Luka Mesin,  
Gabriel Victora, George C. Tsokos, Christian Mawrin,  
Ronald Herbst & Michael C. Carroll

---

In follow-up experiments to this Letter, we have been unable to replicate key aspects of the original results. Most importantly, the findings from behaviour studies and sequencing of microglia isolated from 564Igi autoimmune mice as shown in Figs. 1a, b, d and 3a, b are not substantiated upon further analysis of the original data. The authors therefore wish to retract the Letter. We deeply regret this error and apologize to our scientific colleagues.



## Methods

### Synthesis of the IHA ligand

*O*-tritylhydroxylamine was synthesized as previously described<sup>32</sup>. Chloroacetyl chloride (0.58 ml, 7.3 mmol) was dissolved in 2 ml CH<sub>2</sub>Cl<sub>2</sub> and added dropwise to a suspension of *O*-tritylhydroxylamine (2.0 g, 7.3 mmol) and *N,N*-diisopropylethylamine (2.5 ml, 14.5 mmol) in 15 ml CH<sub>2</sub>Cl<sub>2</sub> at 0 °C. The reaction mixture was gradually warmed to room temperature and stirred at room temperature for an hour. An additional 15 ml CH<sub>2</sub>Cl<sub>2</sub> was added and the reaction was extracted with H<sub>2</sub>O (3 × 30 ml). The CH<sub>2</sub>Cl<sub>2</sub> solution was collected and evaporated to dryness. A solution containing 15 ml of CH<sub>2</sub>Cl<sub>2</sub> with 10% (v/v) trifluoroacetic acid was added and the solution was stirred for 30 min. The crude product was purified by silica gel chromatography using a gradient of 0–100% ethyl acetate in hexanes as the eluent. The product was visualized using a FeCl<sub>3</sub> stain. Yield, 55%. Measured molecular weight (*m/z*): 108.37 [M – H<sup>+</sup>]; calculated: 107.99 [M – H<sup>+</sup>]. <sup>1</sup>H NMR: (300 MHz, DMSO-*d*<sub>6</sub>) δ 10.88 (s, 1H), δ 9.15 (s, 1H), δ 3.93 (s, 2H). <sup>13</sup>C NMR: (500 MHz, DMSO-*d*<sub>6</sub>) δ 162.88, δ 40.45. 2-chloro-*N*-hydroxyacetamide (400 mg, 3.7 mmol) and NaI (2.7 g, 18.3 mmol) were refluxed in 30 ml acetone for 1 h. The reaction mixture was purified by silica gel chromatography with 100% ethyl acetate as the eluent and dried in vacuo. Yield, >90%. Measured molecular weight (*m/z*): 223.85 [M + Na<sup>+</sup>]; calculated: 223.95 [M + Na<sup>+</sup>]. <sup>1</sup>H NMR: (300 MHz, DMSO-*d*<sub>6</sub>) δ 10.81 (s, 1H), δ 9.09 (s, 1H), δ 3.51 (s, 2H). <sup>13</sup>C NMR: (500 MHz, DMSO-*d*<sub>6</sub>) δ 164.83, δ –2.01.

### Protein expression and purification

All constructs (Supplementary Table 1) were derived from the parent pET-20b(+) plasmid containing the *CFMC1* gene via site-directed mutagenesis as previously described<sup>28,33,34</sup>. The appropriate plasmids were transformed into BL21(DE3) *Escherichia coli* cells (New England Biolabs) housing a CCM (cytochrome C maturation) cassette containing a chloramphenicol-resistance marker and expressed as previously described<sup>35</sup> with minor adjustments. Multiple 2.8-l flasks containing 1.5 l of LB medium were shaken at 200 rpm for 12 h at 37 °C and then at 100 rpm for an additional period of around 7 h. Cells were collected by centrifugation (5,000 rpm for 10 min at 4 °C), resuspended in a buffered solution containing 5 mM sodium acetate (NaOAc) (pH 5.0) and 2 mM dithiothreitol (DTT) and lysed via sonication. The pH of the crude lysate was first raised to 10 using NaOH to precipitate cellular contaminants, then reduced to pH 4.5. After centrifugation (12,000 rpm for 20 min at 4 °C), the clarified supernatant was decanted and diluted 15-fold with additional buffer. This solution was applied to a CM sepharose gravity column (GE Healthcare) pre-equilibrated with the aforementioned buffer and subjected to multiple buffer washes before elution using a stepwise-gradient of NaCl (0–0.5 M). Peak elution fractions were combined and concentrated using a 400-ml Amicon Stirred Cell (Millipore) and buffer-exchanged by overnight dialysis against a buffered solution containing 10 mM phosphate (pH 8.0) at 4 °C. Next, the protein was purified via a DuoFlow workstation fitted with a Macroprep High Q-cartridge column (BioRad) and eluted using a linear gradient over 0–0.5 M NaCl. Fractions that exhibited an RZ ratio ( $A_{421}/A_{280}$ ) > 4.4 were pooled, treated with 2 mM EDTA for 1 h, concentrated, and buffer-exchanged into 20 mM Tris(hydroxymethyl)aminomethane (Tris) (pH 7.5) pretreated with Chelex 100 resin (BioRad), via desalting column (Econo-Pac 10DG pre-packed columns, BioRad). Demetallated and purified proteins were concentrated to around 2 mM and stored at 4 °C.

### Protein labelling and post-labelling purification

Purified protein solutions were treated with a 100-fold excess of DTT and placed in an anaerobic Coy chamber for approximately 2 h for slow degassing to remove dissolved oxygen. The fully reduced protein solution was buffer-exchanged into 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) (pH 7.5) via desalting column to

remove DTT, and the concentration of the resulting protein solution was determined spectroscopically (Agilent 8452 spectrophotometer) using the  $\epsilon_{421(\text{red})} = 162,000 \text{ M}^{-1} \text{ cm}^{-1}$  (ref. <sup>34</sup>). Solid iodohydroxamic acid (IHA) was dissolved in 100  $\mu\text{l}$  degassed DMF to generate solutions containing a 15-fold excess IHA per protein monomer, which were then added to protein aliquots and incubated overnight. The HA-functionalized variants were removed from the Coy chamber and separated from unreacted or partially reacted protein via FPLC using a Q-column equilibrated with 10 mM *N*-cyclohexyl-2-aminoethanesulfonic acid (CHES) (pH 9.3) and 2 mM DTT and eluted using a linear gradient over 0–0.5 M NaCl. Protein functionalization was verified using electrospray ionization mass spectrometry (ESI-MS; Extended Data Fig. 1) and the resulting protein solutions were buffer-exchanged into demetallated 20 mM Tris (pH 7.5) via desalting column, concentrated to around 2 mM and stored at 4 °C for further use.

### Redesign of CFMC1 interfaces

To render the CFMC1 protomer competent for the bimetallic design strategy, we first performed the following mutations to remove potential competitive interactions: C67E, H59S and H73N. A negative design strategy was then used to disrupt a noncovalent dimerization interface found in CFMC1, leading to the mutations A34Q and A38Q. We further identified the dearth of protein–protein interactions within the core and periphery of the three-fold axis engulfing the 82 position as a likely contributor to poor cage assembly and crystallization in general. Accordingly, as a means to facilitate cage formation, we adopted Rosetta-prescribed mutations at the following positions: A24T, Q25(T/E), N80K and E81Q.

### Crystallography

Screening and crystallization of all BMC variants were conducted via sitting drop vapour diffusion. In brief, solutions containing 2.1–2.2 mM BMC protomer were mixed with mother liquor (1  $\mu\text{l}$  + 1  $\mu\text{l}$ ) and equilibrated against 200- $\mu\text{l}$  reservoir volumes. Supplementary Table 2 details the experimental conditions for crystal growth. Protein solutions of BMC1 and BMC4 were first incubated with FeSO<sub>4</sub> for 1 h before mixing with ZnCl<sub>2</sub>. Solutions of BMC2 and BMC3 were mixed with FeSO<sub>4</sub> and ZnCl<sub>2</sub> stock solutions and were immediately combined with the mother liquor (to prevent rapid aggregation of the proteins functionalized with two HA units). Crystals for all mutants typically appeared within several hours and were collected within a week of maturation. Crystals were cryoprotected by submersion into perfluoropolyether cryo oil (Hampton Research) for a few seconds and flash-frozen in liquid nitrogen. X-ray diffraction data were collected at 100 K at either the Advanced Light Source (ALS) beamline BL 8.3.1 (using 1.12 Å radiation for BMC3 and 1.33 Å radiation for BMC4) or at the Stanford Synchrotron Radiation Lightsource (SSRL) beamlines 9-2 (using 0.98 Å radiation for BMC2) and 12-2 (using 0.98 Å radiation for BMC1). Data integration was performed using the XDS Program Package, truncated at CC<sub>1/2</sub> > 0.5 (ref. <sup>36</sup>). Datasets of the same structure recorded at different wavelengths were scaled to the highest resolution dataset with XSCALE<sup>37,38</sup>. Phaser-MR<sup>39</sup> was used to carry out molecular replacement with search models based on the CFMC1 monomer (PDB ID: 3M4B) containing the expected side chain mutations (generated in Pymol<sup>40</sup>) but lacking HA. Rigid-body and structure refinement was performed using multiple rounds of Phenix.refine<sup>39</sup>, interspersed with manual model rebuilding and metal/ligand placement with Coot<sup>41</sup>. Restraint files for the Cys-hydroxamic acid conjugates were generated using phenix.eLBOW to maintain the distances Cys-SG–HA-C1 (1.816 Å ± 0.02 Å) and angles Cys-CB–Cys-SG–HA-C1 as well as Cys-SG–HA-C1–HA-C2 (both 109° ± 3°) during refinement. Where necessary, the metal binding geometry of the hydroxamic acids was restrained to the distances Fe–HA-O1 (1.98 Å ± 0.05 Å) and Fe–HA-O2 (2.057 Å ± 0.05 Å) as well as through a planarity constraint for the atoms Fe, HA-O1, HA-O2 and HA-C1 following data from a high-resolution structure of Fe(III)-tris-benzhydroxamate trihydrate<sup>42</sup>. Simulated



# Article

annealing omit maps (metal atoms and side chain ligands) were generated for each metal binding site and model accuracy was assessed critically against these omit maps. Electron density maps were generated using Phenix and all molecular graphics images were produced with either Pymol or the UCSF ChimeraX package from the Computer Graphics Laboratory, University of California, San Francisco<sup>43</sup>.

## Crystallographic metal content analysis

Metal ions, with their relatively high-energy inner electrons, can absorb and resonate with soft X-rays; this leads, among other effects, to differences in the intensity of otherwise centro-symmetric Bragg diffraction peaks used for X-ray crystallography. Density maps calculated from these differences are routinely used to locate and identify metal ions in protein crystals. The magnitude of this anomalous X-ray diffraction varies with the X-ray energy, with stark differences around the energies of the K- and L-shell electrons of the respective elements allowing one to discern between elements at a position in question, if diffraction datasets are measured at the appropriate wavelengths. For a visual analysis of the bound metals, the scaled datasets of different wavelengths were used separately as an input for a single phenix.refine run, each with the final model of the highest resolution dataset. Importantly, only the B-factor or occupancy were allowed to change during refinement, resulting in anomalous difference density maps for each wavelength. Using these maps, isomorphous difference maps from data at wavelengths above and below the respective element K-edges were generated (if applicable) with Phenix and were inspected manually (Extended Data Fig. 5). To gain a more quantitative understanding of the identity of the bound metals for each site, the anomalous difference signal of each dataset was used to generate CCP4 format maps with phenix.mtz2map. The generated maps were used subsequently as inputs to calculate the mean signal in a sphere of 1 Å radius centred on each metal atom with the program MAPMAN (Uppsala Software Factory). For each pair of datasets above and below a metal-absorption edge, the ratio of the anomalous signal above and below the edge for every metal atom was tabulated. The experimental ratio was compared to the theoretical ratio for both Fe and Zn (Extended Data Fig. 5) according to <http://skuld.bmsc.washington.edu/scatter>, as calculated using the Cromer and Liberman approximation. Theoretical ratios were also calculated for hypothetical mixed occupancy Fe/Zn metal sites and compared to experimentally observed values (Supplementary Tables 3–6).

## Protein cage sample preparation

**Self-assembled cages.** All samples were prepared in a low-O<sub>2</sub> atmosphere (Coy glovebox) to minimize undesired oxidation of Fe<sup>2+</sup> ions before self-assembly. Protein solutions containing 20 μM BMC3 or 100 μM BMC4 in 20 mM Tris (pH 8.5) were incubated with either 20 μM FeSO<sub>4</sub> and 60 μM ZnCl<sub>2</sub> for BMC3 or 50 μM FeSO<sub>4</sub> and 200 μM ZnCl<sub>2</sub> for BMC4 for 2–3 h to yield the metallated cages. We note that the addition of FeSO<sub>4</sub> was followed by a small but observable change in the colour of the solution from red to pink, attributed to a shift of the haem Soret band to longer wavelengths, which suggested reduction of the haem by the ferrous ions and generation of ferric ions in close proximity to HA group(s). The final BMC3 solutions were then concentrated sevenfold before overnight incubation to improve the total cage yield. After self-assembly, the resulting solutions were diluted back to their original concentrations with the self-assembly buffer before characterization.

**Dissolved crystals.** Fe:Zn:BMC1 and Fe:Zn:BMC2 crystals were dissolved using buffer containing 100 mM HEPES (pH 7.5), 200 mM MgCl<sub>2</sub> and 800 μM ZnCl<sub>2</sub>. Mature crystals were removed from their pedestal droplet, briefly submerged in fresh buffer to remove uncrystallized protein and surface-bound precipitates, and transferred into a new sitting drop crystallization well containing 8 μl buffer solution. The crystals were physically crushed with a small metal scalpel and vigorously pipetted until a large portion of the crystals dissolved. Undissolved crystals

were removed by centrifugation (10,000 rpm for 5 min at 25 °C), yielding a light-red supernatant and dark-red precipitate.

## Negative-stain transmission electron microscopy

A 4-μl droplet of BMC cages (either self-assembled or from dissolved crystals) was deposited onto formvar/carbon-coated Cu grids (Ted Pella) (pretreated by negative-mode glow discharge up to 15 min beforehand) and allowed to bind for 5 min. The grids were then washed with 50 μl MilliQ water, blotted using Whatman filter paper and stained using 2% uranyl acetate solution in water and blotted again. Grids were imaged using a FEI Sphera transmission electron microscope operating at 200 keV, equipped with an LaB<sub>6</sub> filament and a Gatan 4K CCD camera. Micrographs were collected using objective-lens underfocus settings ranging from 250 nm to 2 μm and analysed using Fiji (<http://fiji.sc/Fiji>).

## Oligomerization state determination using AUC

Sedimentation velocity measurements were performed at 41,000 rpm and 25 °C using an XL-1 analytical ultracentrifuge (Beckman Coulter) equipped with an AN-60 Ti rotor. Data processing was performed using Sedfit<sup>44</sup> with the following parameters as calculated using SEDNTERP: viscosity: 0.01000 poise, density: 0.9988 g/ml (self-assembled samples) or viscosity: 0.0113191 poise, density: 1.0196 g/ml (dissolved crystals), and a partial specific volume of 0.7313 ml/g for all samples. All reported results correspond to a confidence level of 0.95.

## Preparation of samples involving crystal dissolution

Dissolved crystal samples (BMC1 and BMC2), prepared as described above at ambient conditions, were diluted to 350 μl with 10 mM HEPES (pH 7.5), 200 mM MgCl<sub>2</sub> and 800 μM ZnCl<sub>2</sub>. The solution was clarified via brief centrifugation in order to remove crystal debris and the supernatant was placed inside the cells.

## Calculation of BMC void volumes

Structures of complete cage assemblies for BMC2, BMC3 and BMC4 were generated via the application of crystallographic symmetry operations to the fully refined asymmetric unit of each construct. These coordinates were recentred at the origin and stripped of waters, hydrogens, alternative conformations and crystallization reagents (PEG-400). Volumetric maps and volumes for the internal cavity of each cage were calculated using VOIDOO<sup>45</sup>, and are reported as the solvent-accessible volume for a 1.4 Å rolling probe on a 0.25 Å grid spacing for all constructs. The cavity volumes using these parameters were determined to be approximately 32,700 Å<sup>3</sup> (BMC2), 32,700 Å<sup>3</sup> (BMC3) and 7,900 Å<sup>3</sup> (BMC4).

## Solution self-assembly, disassembly and thermal stability of BMC3 and BMC4

Assembled samples were prepared as described above and placed inside the AUC measurement cells anaerobically (20 μM BMC3 and 100 μM BMC4). Disassembly of the cages via metal-ion removal was performed by treating the protein cages with 2 mM EDTA for 1 h. Redox-controlled disassembly of the protein commenced by the addition of either 5 mM sodium dithionite or 5 mM sodium ascorbate to the cage solution anaerobically and subsequent incubation of the samples at around 22 °C for 16 h. Samples were then loaded into the AUC measurement cell.

For thermal stability measurements, samples were placed in a thermoregulated chamber pre-equilibrated at the appropriate temperature for 2 h, and subsequently removed from the chamber and equilibrated at room temperature for 30 min before AUC analysis. Circular dichroism (CD) spectra were measured using an Aviv 215 spectrometer. CD measurements were performed using 10 μM protein in a buffered solution containing 20 mM Tris (pH 8.5). Thermal melts were measured at 222 nm at a 1 nm slit width, scanning at 1-nm intervals with a 1-s integration time. Measurements were taken from 25 °C to 85 °C at 2-degree intervals

with a 2 min equilibration at each temperature. Unfolding data were fit to a two-state model with van't Hoff's enthalpy using the CalFitter web server<sup>46</sup>.

### Cryo-EM sample preparation

Self-assembled BMC3 cages were removed from the anaerobic Coy chamber immediately before grid preparation. A 3.5- $\mu$ l aliquot of self-assembled BMC3 cages was dropped onto holey carbon grids (Electron Microscopy Sciences, Quantifoil R1.2/1.3 holey carbon on 300 mesh copper) that had been freshly glow-discharged for 30 s. The initial application of the sample was side blotted manually with Whatman No. 1 filter paper immediately followed by a secondary application of a 3.5- $\mu$ l aliquot, blotted for 3.5 s and plunge-frozen in liquid ethane cooled by liquid nitrogen using a Vitrobot Mark IV (FEI).

### Cryo-EM data acquisition and image processing

Samples were imaged on a Titan Krios G3 transmission electron microscope (FEI) operating at 300 kV equipped with a K2 Summit direct electron detector (Gatan) and a GIF Quantum energy filter. The slit-width of the energy filter was set to 10 eV. Movies were collected at a magnification of 165,000 $\times$  in EFTEM mode giving a physical pixel size of 0.84  $\text{\AA}$ /pixel. In total, 4,672 movie stacks (50 frames/movie) were collected using a 10 s exposure at a dose rate of 1.2  $e^-/\text{\AA}^2$  per frame for a total electron dose of 60  $e^-/\text{\AA}^2$  per movie. Objective-lens underfocus settings varied between 0.6  $\mu$ m and 1.6  $\mu$ m. Data collection was performed using software EPU (FEI). All image processing was performed in the Relion-3.0 pipeline<sup>47</sup>. Motion correction and dose weighting were performed using MotionCor2<sup>48</sup>, and defocus values were estimated with Gctf<sup>49</sup> using a pixel size of 0.8  $\text{\AA}$ /pixel. A total of 3,513 movie stacks were selected following motion correction and CTF estimation, and 805,156 particles were auto-picked using RELION-3.0. Particle images were extracted and binned by 2 (1.6  $\text{\AA}$ /pixel, 100 pixel box size) and subjected to two-dimensional (2D) classification. A total of 444,247 particles were selected corresponding to good 2D class averages and subjected to three-dimensional (3D) classification imposing *T* symmetry and using an initial model generated from a subset of the particles. A total of 129,653 particles were chosen from a 3D class showing strong secondary-structural elements and subjected to 3D auto-refinement with *T* symmetry. The particles were re-centred and re-extracted to their original pixel size of 0.8  $\text{\AA}$ /pixel. These particles were subjected to 3D auto-refinement with *T* symmetry and the yield map was then postprocessed towards 2.6  $\text{\AA}$  resolution based on the gold-standard Fourier shell correlation (FSC) 0.143 criterion. The pixel size of the map was manually adjusted using Relion image handler to match the physical pixel size of the images. Local resolution was calculated in Relion 3.0 using ResMap<sup>50</sup>.

### Model building and refinement

The BMC3 crystal structure (PDB ID: 6OT7) stripped of hydrogens and waters was used as an initial model and manually docked into the cryo-EM density using UCSF Chimera<sup>51</sup>. The structural model was subject to real space refinement in Phenix against the cryo-EM map with geometry restraints for the Fe-binding sites and molecular coordinates for the Cys-HA ligand. The atomic model was manually improved using Coot. Tightly bound waters were identified based on clear density in the EM density map. Whereas the structural flexibility of the hydroxamate sites manifested in poor electron density, the twofold interface was much more rigid and unambiguous density was observed for Zn-binding. A tryptophan at the 66 position, which had shown high-temperature factors in the BMC3 crystal structure, was identified in multiple conformations in the EM density map. The final model was subjected to real space refinement using Phenix<sup>39</sup> and evaluated using MolProbity<sup>52</sup>. All molecular graphics images were rendered in PyMol or UCSF ChimeraX.

### Encapsulation of rhodamine in BMC3 cages

BMC3 cages were self-assembled in a low- $O_2$  atmosphere in the presence of rhodamine for the passive encapsulation of the dye. Solutions containing 20  $\mu$ M BMC3 were incubated with 20  $\mu$ M  $\text{FeSO}_4$ , 60  $\mu$ M  $\text{ZnCl}_2$  and 2 mM rhodamine. A control sample was prepared in the absence of added metal ions (20  $\mu$ M BMC3 incubated with 2 mM rhodamine). Samples were incubated for 2–3 h and concentrated sevenfold before overnight incubation. Protein solutions were buffer exchanged on a PD-10 desalting column using a buffer containing 20 mM Tris (pH 8.5) (with 5  $\mu$ M  $\text{FeSO}_4$  and 10  $\mu$ M  $\text{ZnCl}_2$  supplemented for solutions already containing metal ions) to separate unassociated dye from protein. Cage solutions were split in two: one half was treated with 1 mM EDTA and incubated for 2 h before washing. All protein solutions were additionally washed three times using a centrifugal filter to completely remove any remaining free rhodamine.

Fluorescence measurements were performed using 6  $\mu$ M protein solutions after the previously mentioned wash steps. For each sample, an excitation wavelength of 555 nm with a 2 nm slit width was used and emission was measured between 560 and 650 nm with a 2 nm slit width and 0.2 s integration time. For the time-course experiments, cages encapsulating rhodamine were washed three times after 4 days and after 7 days and diluted to 6  $\mu$ M before fluorescence measurements. AUC measurements were performed at the  $\lambda_{\text{max}}$  of the cytochrome (415 nm) and at the  $\lambda_{\text{max}}$  of rhodamine (555 nm) to assess whether there was a sufficiently large rhodamine signal associated with BMC3 cages. Ultra-violet–visible light (UV-vis) absorbance measurements were performed on each solution to measure the protein and rhodamine concentrations. Difference spectra were taken between each rhodamine-incubated sample and BMC3 protomer to eliminate any background signal.

### Statistics and reproducibility

All reported samples represent technical replicates. The ns-TEM micrograph of BMC2 cages after 3D crystal dissolution (Fig. 2a) is representative of experiments repeated independently four times. AUC experiments for BMC2 (Fig. 2b) were performed in duplicate. Self-assembly of BMC3 cages and subsequent AUC characterization (Fig. 3a) were performed the following number of times: BMC3 protomer ( $n = 2$ ),  $+\text{Fe}^{2+}$  ( $n = 4$ ),  $+\text{Zn}^{2+}$  ( $n = 4$ ),  $+\text{Fe}^{2+}$ ,  $+\text{Zn}^{2+}$  ( $n = 6$ ). Self-assembly of BMC4 cages and subsequent AUC characterization (Fig. 4a) was performed the following number of times: BMC4 protomer ( $n = 2$ ),  $+\text{Fe}^{2+}$  ( $n = 1$ ),  $+\text{Zn}^{2+}$  ( $n = 1$ ),  $+\text{Fe}^{2+}$ ,  $+\text{Zn}^{2+}$  ( $n = 5$ ). Mass spectra (Extended Data Fig. 1c–f) were collected in duplicate for native and HA-labelled proteins; AUC experiments were performed in duplicate. TEM characterization of BMC constructs (Extended Data Fig. 3) were performed the following number of times: dissolved BMC1 crystals ( $n = 1$ ), dissolved BMC2 crystals ( $n = 4$ ), BMC2 + EDTA ( $n = 2$ ), self-assembled BMC3 cages ( $n = 5$ ), BMC3 + EDTA ( $n = 4$ ). AUC experiments following the incubation of BMC3 with first-row transition metals (Extended Data Fig. 6a) were performed in duplicate. Self-assembly of BMC3 in the presence of  $\text{Fe}(\text{acetylacetonate})_3$  (Extended Data Fig. 6b) was performed in duplicate. BMC3 cage disassembly in the presence of EDTA (Extended Data Fig. 6c) was performed in triplicate. AUC characterization of BMC variants after equilibration at different temperatures (Extended Data Fig. 6d) was performed the following number of times: BMC3 at 50  $^\circ\text{C}$  ( $n = 2$ ), BMC3 at 70  $^\circ\text{C}$  ( $n = 2$ ), BMC4 at 50  $^\circ\text{C}$  ( $n = 3$ ), BMC4 at 70  $^\circ\text{C}$  ( $n = 3$ ), BMC4 at 90  $^\circ\text{C}$  ( $n = 3$ ). Thermal unfolding of BMC variants as measured by CD spectroscopy (Extended Data Fig. 6d) was performed in duplicate. Treatment of BMC3 cages with chemical reductants (Extended Data Fig. 6e) was performed in duplicate. Cryo-EM characterization of BMC3 cages was performed after collecting 4,672 movie stacks. Extended Data Figure 7a shows a representative micrograph and three representative 2D class averages. Fluorescence characterization of BMC3 samples incubated with rhodamine were performed (Extended Data Fig. 8a) in triplicate. AUC characterization

of BMC3 cages encapsulating rhodamine (Extended Data Fig. 8b) was performed in duplicate. UV-vis characterization of BMC3 samples incubated with rhodamine (Extended Data Fig. 8c, d) was performed in triplicate. Repeated fluorescence characterization of a solution containing BMC3 cages encapsulating rhodamine (Extended Data Fig. 8e) was performed in duplicate.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The principal data supporting the findings of this work are available within the figures and the Supplementary Information. Additional data that support the findings of this study are available from the corresponding author on request. Structural data obtained by X-ray crystallography and cryo-EM have been deposited into the RCSB PDB and EMDB data banks with the following accession codes: 6OT4 (BMC2), 6OT7 (BMC3), 6OT8 (BMC4), 6OT9 (BMC1) and 6OVH (BMC3 cryo-EM) in the PDB or EMD-20212 at The Electron Microscopy Data Bank.

32. Michalak, K., Wicha, J. & Wójcik, J. Studies towards dynamic kinetic resolution of 4-hydroxy-2-methylcyclopent-2-en-1-one and its *E*-*O*-trityloxime. *Tetrahedron* **72**, 4813–4820 (2016).
33. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).
34. Faraone-Mennella, J., Tezcan, F. A., Gray, H. B. & Winkler, J. R. Stability and folding kinetics of structurally characterized cytochrome *c*-*b*<sub>562</sub>. *Biochemistry* **45**, 10504–10511 (2006).
35. Bailey, J. B., Subramanian, R. H., Churchfield, L. A. & Tezcan, F. A. in *Methods in Enzymology* Vol. 580 (ed. Pecoraro, V. L.) 223–250 (Academic, 2016).
36. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).
37. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66**, 133–144 (2010).
38. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
39. Terwilliger, T. C. et al. Phenix—a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
40. Schrodinger, LLC. The PyMOL Molecular Graphics System. version 1.3 (2010).
41. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
42. Lindner, H. J. & Gottlicher, S. Die Kristall- und Molekülstruktur des Eisen(III)-benzhydroxamat-Trihydrates. *Acta Crystallogr. B* **25**, 832–842 (1969).

43. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
44. Schuck, P. A model for sedimentation in inhomogeneous media. I. Dynamic density gradients from sedimenting co-solutes. *Biophys. Chem.* **108**, 187–200 (2004).
45. Kleywegt, G. J. & Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D* **50**, 178–185 (1994).
46. Mazurenko, S. et al. CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res.* **46**, W344–W349 (2018).
47. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
48. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
49. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
50. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
51. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
52. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).

**Acknowledgements** This work was supported by the US Department of Energy (Division of Materials Sciences, Office of Basic Energy Sciences; DE-SC0003844; for the design strategy, EM imaging and analysis, and biochemical analysis) and by the National Science Foundation (Division of Materials Research; DMR-1602537; for crystallographic analysis). E.G. acknowledges support by an EMBO Long-Term Postdoctoral Fellowship (ALTF 1336-2015). J.E. acknowledges support by a DFG Research Fellowship (DFG 393131496). R.H.S. was supported by the National Institute of Health Chemical Biology Interfaces Training Grant UC San Diego (T32GM112584). We acknowledge the use of the UCSD Cryo-EM Facility, which is supported by NIH grants to T.S.B. and a gift from the Agouron Institute to UCSD. Crystallographic data were collected either at Stanford Synchrotron Radiation Lightsource (SSRL) or at the Lawrence Berkeley National Laboratory on behalf of the Department of Energy.

**Author contributions** E.G. conceived the project, and designed and performed most experiments. R.H.S. and R.G.A. performed and processed the ns-TEM data and performed structural modelling and analysis. R.H.S. conducted encapsulation experiments. J.E. performed crystallographic analysis. J.B.B. and J.A.C. synthesized the IHA ligand. X.Y., T.B. and T.S.B. performed the cryo-EM data collection and processing. F.A.T. conceived and directed the project and wrote the manuscript with contributions from all authors.

**Competing interests** The authors declare no competing interests.

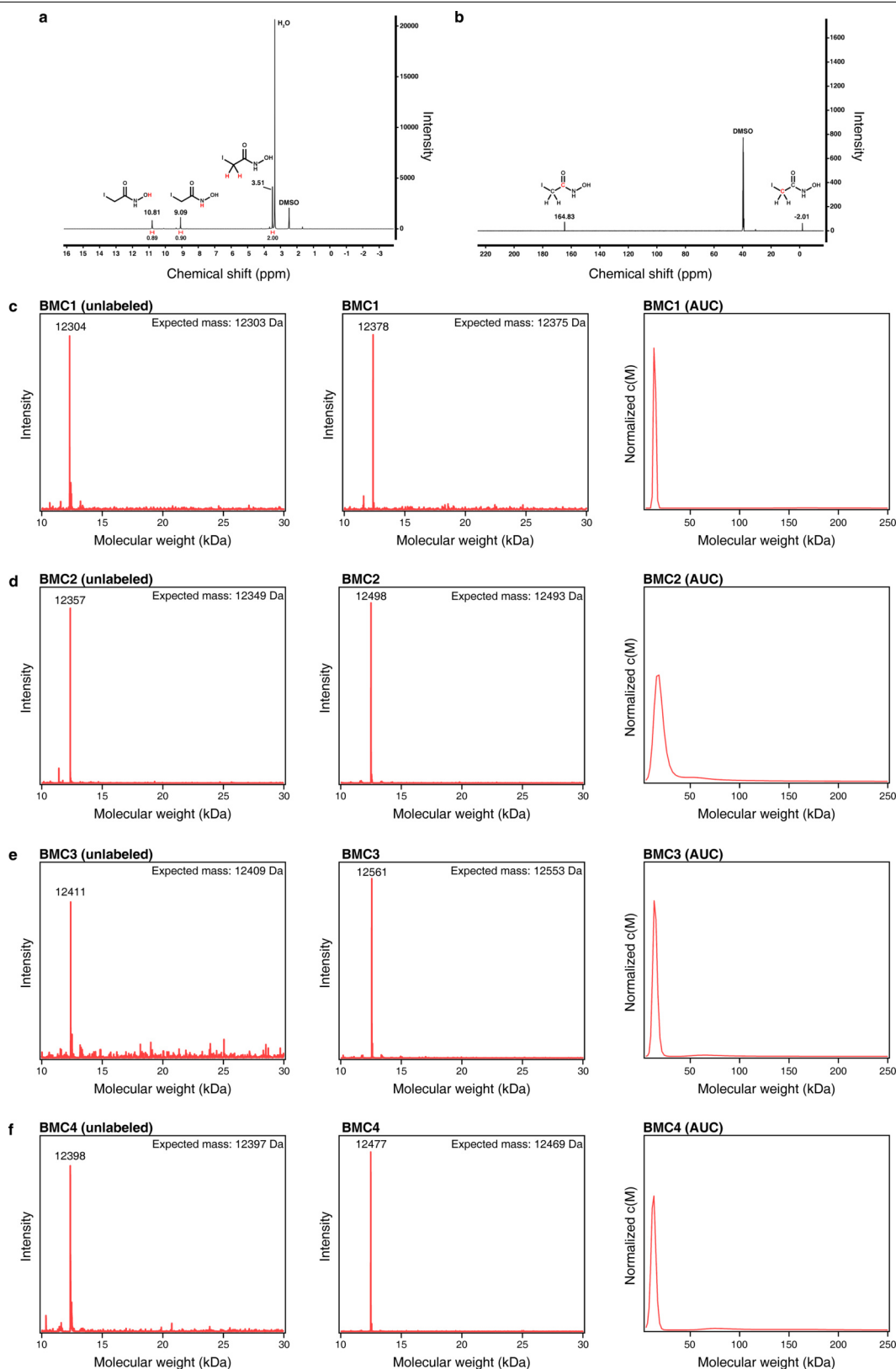
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1928-2>.

**Correspondence and requests for materials** should be addressed to F.A.T.

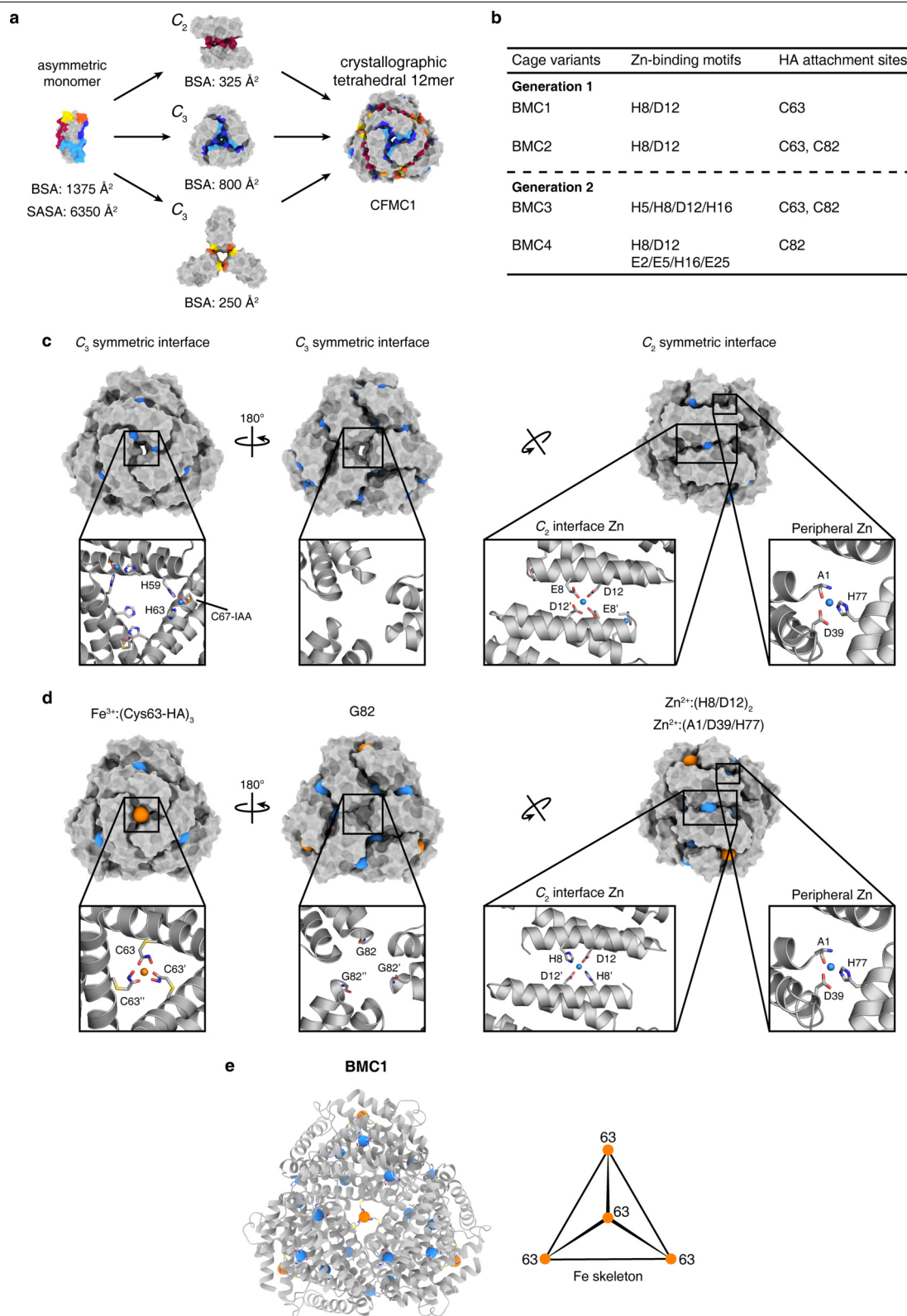
**Peer review information** *Nature* thanks Jack Johnson, Todd Yeates and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Characterization of the IHA ligand and the BMC constructs.** **a, b**, NMR spectra of *N*-hydroxy-2-iodoacetamide in  $\text{DMSO-}d_6$  for  $^1\text{H}$  (**a**) and  $^{13}\text{C}$  (**b**). **c–f**, ESI-MS of as-isolated and HA-functionalized BMC constructs, and AUC profiles of HA-functionalized protomers for BMC1

(**c**), BMC2 (**d**), BMC3 (**e**) and BMC4 (**f**). The calculated masses for each unlabelled protein are determined by summing the mass of the polypeptide sequence and the *c*-type haem (618 Da) covalently linked to the cytochrome.

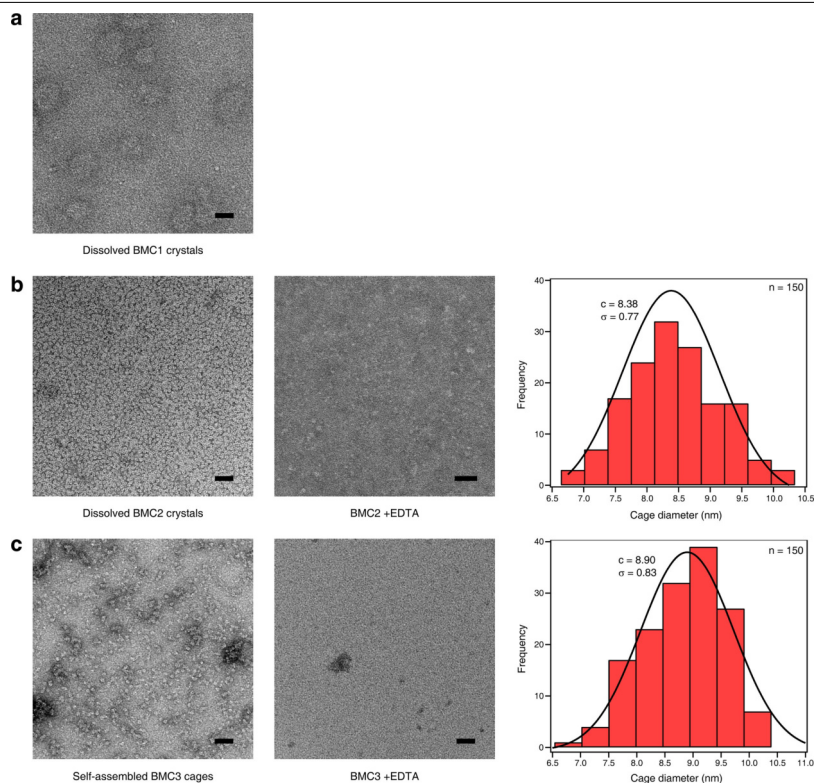


**Extended Data Fig. 2 | Structural comparison of CFMC1 and BMC1 cages.**

**a**, The symmetric substructures of the CFMC1 dodecameric unit and its per-protomer SASA and BSA values. Associative surfaces on the protomers are coloured red for homologous interactions and red/orange or blue/cyan for heterologous interactions (right). **b**, Summary of engineered metal-coordination motifs for BMC constructs (see Supplementary Table 1 for all

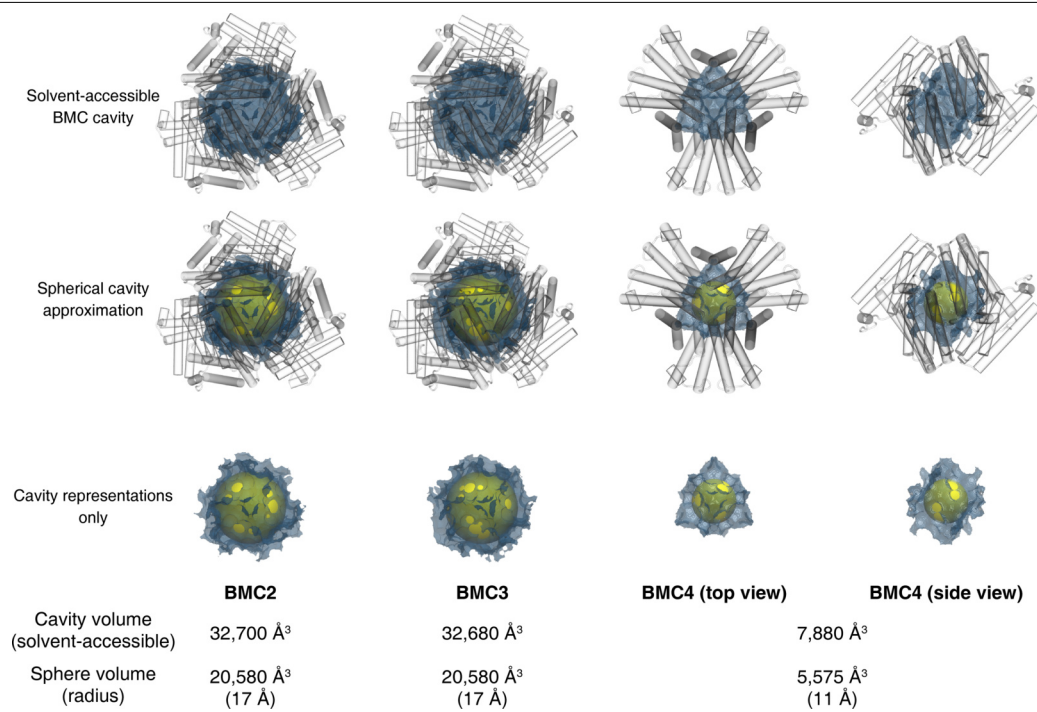
mutations). **c**, **d**, Comparison of  $C_2$  and  $C_3$  symmetric interfaces and corresponding metal binding sites for CFMC1 (**c**) and BMC1 (**d**). Full cages are shown as surfaces; insets show details of each interface. Fe and Zn ions are represented as orange and teal spheres, respectively. **e**, Cartoon representation of a full-size BMC1 cage with all metal ions shown as spheres. PDB ID: 3M4B (CFMC1), 6OT9 (BMC1).



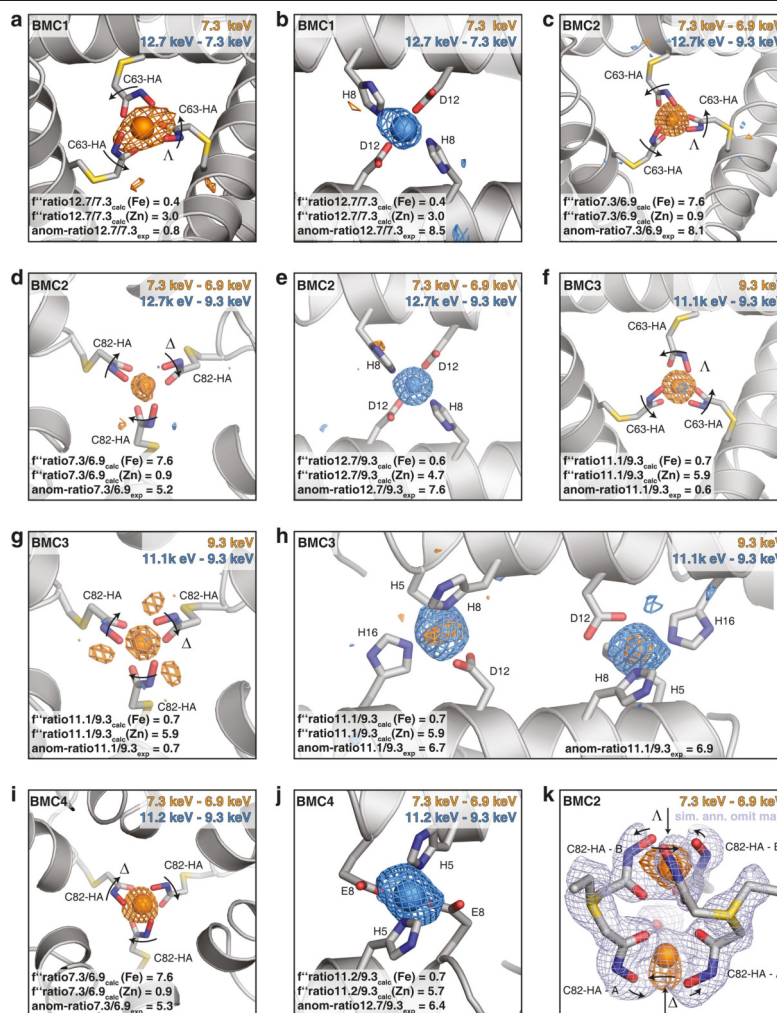


**Extended Data Fig. 3 | ns-TEM characterization of BMC constructs.** **a, b**, Dissolved Fe:Zn:BMC1 (**a**) and Fe:Zn:BMC2 crystals (**b**) in a buffer containing 100 mM HEPES (pH 7.5), 200 mM MgCl<sub>2</sub> and 800  $\mu$ M ZnCl<sub>2</sub>. **c**, Self-assembled Fe:Zn:BMC3 cages in a buffer containing 20 mM Tris (pH 8.5),

20  $\mu$ M FeSO<sub>4</sub> and 60  $\mu$ M ZnCl<sub>2</sub>. Histograms in **b, c** reflect the size distributions of Fe:Zn:BMC2 and Fe:Zn:BMC3 cage diameters as measured from ns-TEM images. Gaussian fits to both distributions are drawn as solid lines along with their centres and standard deviations reported. Scale bars, 50 nm.

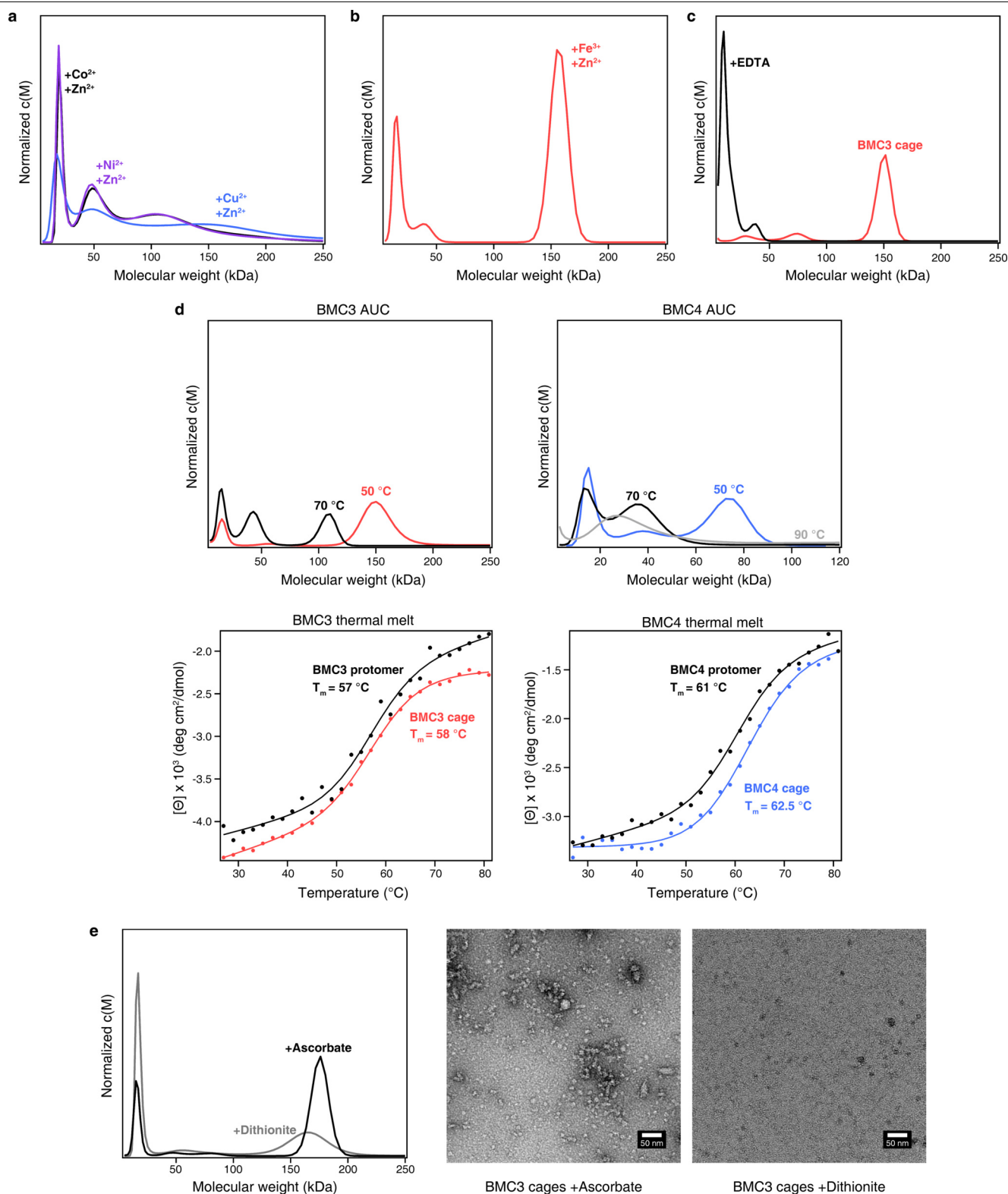


**Extended Data Fig. 4 | Cavity volumes of BMC cages.** Solvent-accessible cavity volumes within BMC cages as calculated by a 1.4 Å rolling probe are shown visually as blue meshes and reported numerically below. Spherical cavities, shown as yellow spheres in Figs. 2, 4, are reproduced for comparison to the calculated volumes. BMC proteins are represented as transparent cylinders.



**Extended Data Fig. 5 | Anomalous densities of engineered metal binding sites and conformational flexibility of Cys82-HA site. a–j,** Cartoon and stick representations of the symmetric interfaces of BMC1 (a, b), BMC2 (c–e), BMC3 (f–h) and BMC4 (i, j) showing the engineered metal binding sites with the C63-HA ligands (a, c, f), C82-HA ligands (d, g, i) and Zn binding sites (b, e, h, j). The difference in the anomalous signal between pairs of datasets above and below the K-shell energy of Zn and Fe, respectively, are depicted as blue or orange meshes. A strong signal illustrates a strong change in anomalous signal across the respective edge, in turn suggesting the presence of the respective metal. The top right corner of each panel indicates the energies of the datasets used for the map of the respective colour. All anomalous difference maps are contoured at  $3\sigma$ . As datasets around the Fe-edge were not available for BMC1 and BMC3 (necessitating calculations using anomalous difference density of singular datasets), the calculated  $f''$  values for Zn at 7.3 and 9.3 keV are 0.82 and

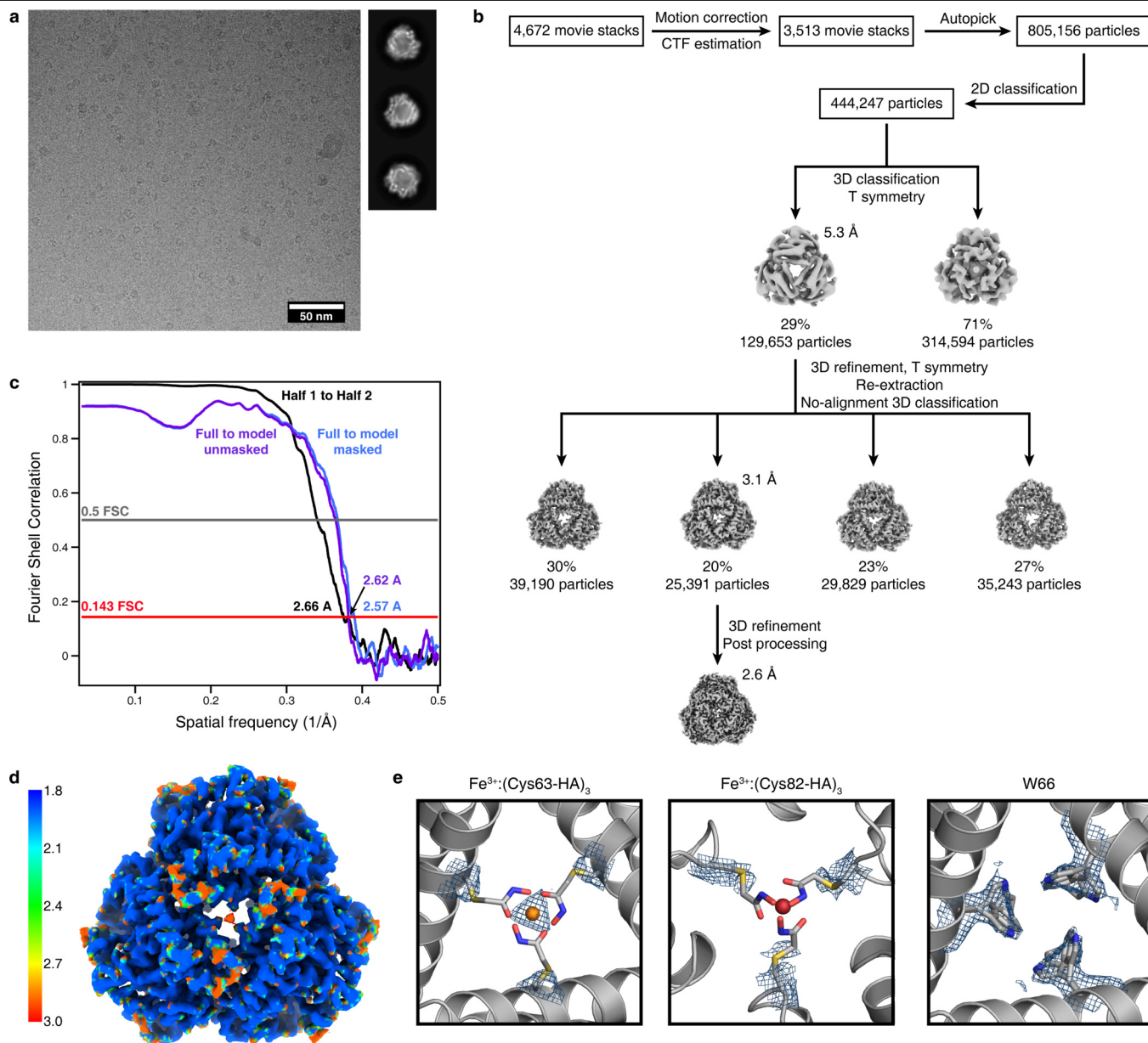
0.52 (that is, non-zero) and thus some residual anomalous signal of the lower energy maps around the Zn atoms is expected to result even from strictly selective Zn loading. For a more quantitative analysis of the nature of the bound metal, ratios of the anomalous signal to the expected values (bottom left corner of each panel) were calculated as described in the Methods. k, Stick representation of the BMC2 Cys82-HA binding site in both alternative conformations with the anomalous difference density over the Fe-edge shown as orange mesh and a simulated annealing omit map (omitting all C82-HA atoms and Fe) of the normal electron density as light blue mesh contoured at  $2\sigma$ . For all Cys-HA binding sites, arrows indicate the handedness of the binding site as  $\Delta$  (right handed) or  $\Lambda$  (left handed). The reversion of handedness in k with the respective view angle is indicated by arrows. Colour code for atoms in all panels: Fe in orange, Zn in blue, S in yellow, O in red and N in dark blue.



**Extended Data Fig. 6 | Solution characterization of self-assembled BMC3 and BMC4 cages.** **a–c**, The oligomerization state of BMC3 cages as monitored by AUC measurements following incubation with various first-row transition metal ions (**a**), incubation with Zn<sup>2+</sup> and Fe<sup>3+</sup> (Fe(acetylacetonate)<sub>3</sub>) (**b**) and disassembly via sequestration of metal ions by EDTA (**c**). **d**, AUC profiles of BMC variants after equilibration for 2 h at the indicated temperatures (top). Thermal

unfolding of BMC variants as measured by circular dichroism spectroscopy at 222 nm (bottom). **e**, AUC profiles of BMC following treatment with chemical reductants of different reduction potentials (left). ns-TEM micrographs (middle and right) are shown for cage samples incubated with the corresponding chemical reductants.

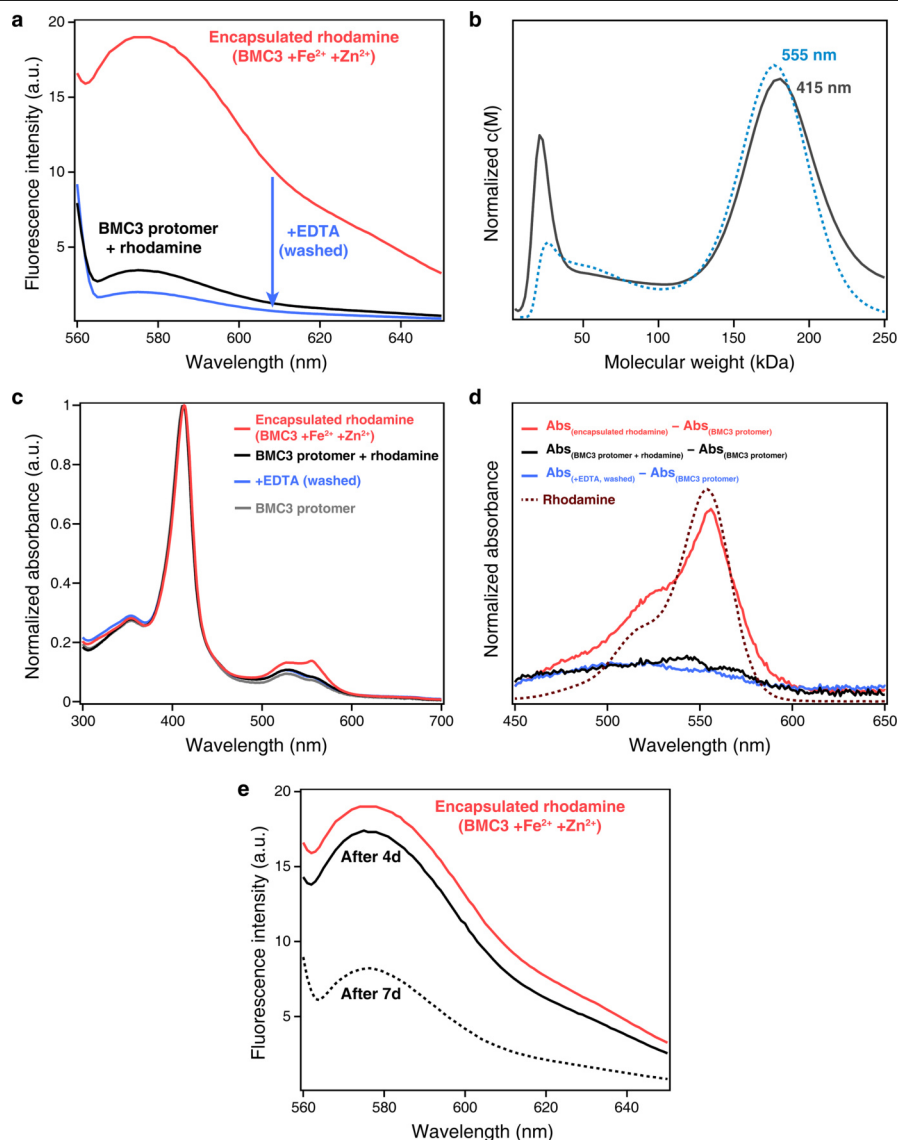




**Extended Data Fig. 7 | Cryo-EM analysis of BMC3 cages. a**, Representative cryo-EM micrograph and 2D class averages. **b**, Flowchart detailing image processing from collected movie stacks to final map. Additional details can be found in the Methods. **c**, FSC curves calculated between the half-maps (black line), atomic model to the unmasked full map (purple line) and atomic model to

the masked full map (blue line). Resolution values are indicated at the gold-standard FSC 0.143 criterion. **d**, Local resolution estimates of the final reconstruction calculated using ResMap. **e**, Electron density shown at BMC3  $C_3$  interfaces highlighting poorly resolved density (reflecting high flexibility) at hydroxamate sites and multiple conformations of W66.





# Extended Data Fig. 8 | Encapsulation of rhodamine inside BMC3 cages.

**a**, Fluorescence characterization of BMC3 samples incubated with rhodamine. Cages encapsulating rhodamine were treated with EDTA and washed before measuring fluorescence intensity. **b**, AUC profiles of cages encapsulating rhodamine monitored at the haem Soret absorption maximum ( $\lambda_{\text{max}} = 415 \text{ nm}$ ) and rhodamine absorption maximum ( $\lambda_{\text{max}} = 555 \text{ nm}$ ). **c**, UV-vis characterization

of BMC3 samples incubated with rhodamine. **d**, Difference spectra of BMC3 samples and BMC3 protomer shown in **c**. Free rhodamine dissolved in buffer is shown as dark-red dashes. **e**, Repeated fluorescence characterization of a solution containing BMC3 cages encapsulating rhodamine over several days. The sample was washed three times before each fluorescence measurement.

Extended Data Table 1 | X-ray data collection, processing and refinement statistics

	BMC1	BMC2	BMC3	BMC4
Data collection				
Space group	R 3 2	R 3 2	R 3 2	P 6 <sub>3</sub> 2 2
Cell dimensions				
<i>a</i> , <i>b</i> , <i>c</i> (Å)	125.6, 125.6, 166.4	126.1, 126.1, 168.2	126.7, 126.7, 167.8	87, 87, 63.3
$\alpha$ , $\beta$ , $\gamma$ (°)	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)	39.92 – 2.40 (2.46 – 2.40)	39.25 – 1.40 (1.44 – 1.40)	91.83 – 1.85 (1.90 – 1.85)	48.47 – 1.50 (1.54 – 1.50)
No. Reflections Observed	393414 (28956)	1863896 (88779)	885564 (66924)	734764 (18399)
No. Reflections Unique	38245 (2875)	195424 (14465)	85783 (6342)	37652 (1334)
<i>R</i> <sub>merge</sub>	0.185 (2.986)	0.056 (1.830)	0.096 (2.377)	0.053 (1.936)
<i>I</i> / $\sigma$ <i>I</i>	8.1 (0.96)	20.3 (0.84)	13.7 (1.04)	26.6 (1.23)
<i>CC</i> 1/2	0.998 (0.501)	1.000 (0.410)	0.999 (0.468)	1.000 (0.480)
Completeness (%)	99.81 (100.00)	99.37 (98.9)	100.00 (100.00)	87.66 (42.2)
Redundancy	10.29 (10.01)	9.54 (6.138)	10.32 (10.54)	19.51 (13.94)
Wilson B (Å) <sup>2</sup>	52	19	36	29
Refinement				
Resolution (Å)	36.37 – 2.40 (2.43 – 2.40)	31.54 - 1.40 (1.42 - 1.40)	91.83 - 1.85 (1.88 - 1.85)	48.47 - 1.50 (1.53 - 1.50)
No. reflections	38194 (1286)	195270 (8377)	85755 (3974)	37642 (810)
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub>	0.2174/0.2718	0.1659/0.1909	0.1826/0.2106	0.1900/0.2181
No. atoms				
Protein	3292	3502	3459	925
Ligand/ion	200	244	240	58
Water	19	786	342	124
<i>B</i> -factors (Å) <sup>2</sup>				
Protein	70	24	41	42
Ligand/ion	68	22	43	35
Water	63	37	46	49
R.m.s. deviations				
Bond lengths (Å)	0.010	0.011	0.009	0.013
Bond angles (°)	1.29	1.33	1.15	1.33
Clashscore	7	5	7	10
Ramachandran favored (%)	100	100	100	97
Ramachandran allowed (%)	0	0	0	3
Ramachandran outliers (%)	0	0	0	0
Rotamer outliers (%)	1	2	2	1

Numbers in parentheses correspond to the highest-resolution shell.

Extended Data Table 2 | Cryo-EM data collection, processing, and refinement statistics

BMC3 (EMDB ID: EMD-20212) (PDB ID: 6OVH)	
Data collection and processing	
Magnification	165,000x
Voltage (kV)	300
Electron dose (e <sup>-</sup> /Å <sup>2</sup> )	60
Exposure rate (e <sup>-</sup> / Å <sup>2</sup> /s)	6
Defocus range (µm)	0.84
Pixel size (Å)	0.3 – 2.7
Symmetry imposed	T
Total extracted particles (no.)	805,156
Final refined particles (no.)	25,391
Map resolution (Å)	2.57
FSC 0.143 (unmasked/masked)	2.62/2.57
Map resolution range (Å)	∞ – 2.57
Applied B-factor (Å <sup>2</sup> )	-79
Refinement	
Initial model used (PDB code)	6OT7
Model resolution (Å)	2.57
FSC 0.5 (unmasked/masked)	2.74/2.72
FSC 0.143 (unmasked/masked)	2.62/2.57
Model resolution range (Å)	∞ – 2.57
Map sharpening B factor (Å <sup>2</sup> )	-79
Model composition	
Non-hydrogen atoms	10895
Protein residues	1272
Ligand/ion	68
Water	171
B-factors (Å) <sup>2</sup>	
Protein	51
Ligand/ion	59
Water	48
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	0.876
Validation	
MolProbity score	1.33
Clashscore	1.15
Rotamer outliers (%)	4.55
Ramachandran plot	
Favored (%)	99.04
Allowed (%)	0.96
Disallowed (%)	0.00

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Confirmed   |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted<br><i>Give P values as exact values whenever suitable.</i>                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated   |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection	X-ray crystallography: Custom software at SSRL and ALS Electron Microscopy: EPU (FEI)
Data analysis	X-ray crystallography: XDS ver. June 1, 2017 (BUILT=20170615) and XSCALE ver. Jun 17, 2015 (BUILT=20150617) (data integration/scaling), Phaser-MR (molecular replacement), Phenix v.1.13-2998 (model building/refinement), Coot v.0.8.6.1 (visualization, real time refinement), Pymol version 1.3 (molecular graphics), ChimeraX v.0.9 (molecular visualization), VMD v.1.9.3 (molecular visualization), Voidoo v.3.3.4 (cavity measurements), Mapman v.7.8.5 (manipulation and analysis of electron-density map), Fiji (image processing), MotionCor2 v.1.2.1 (motion correction), Gctf v.1.06 (defocus value estimation), Relion v.3.0 (particle picking, classification, refinement), ResMap v.1.1.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The various sample sizes (n) for the different experiments was selected as no significant difference was observed between technical replicates
Data exclusions	No data was excluded
Replication	All results from the experiments were successfully replicated.
Randomization	No randomization was performed as it is not relevant to the current study.
Blinding	No blinding was performed as it is not relevant to the current study.

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
-------------------	--



Research sample	<i>Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i>, all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.</i>
Sampling strategy	<i>Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.</i>
Data collection	<i>Describe the data collection procedure, including who recorded the data and how.</i>
Timing and spatial scale	<i>Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken</i>
Data exclusions	<i>If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Reproducibility	<i>Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.</i>
Randomization	<i>Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.</i>
Blinding	<i>Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.</i>
Did the study involve field work? <input type="checkbox"/> Yes <input type="checkbox"/> No	

## Field work, collection and transport

Field conditions	<i>Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).</i>
Location	<i>State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).</i>
Access and import/export	<i>Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).</i>
Disturbance	<i>Describe any disturbance caused by the study and how it was minimized.</i>

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	<i>Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Validation	<i>Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.</i>

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	<i>State the source of each cell line used.</i>
Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

## Palaeontology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	<i>For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.</i>
Wild animals	<i>Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<i>Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural &amp; social sciences study design questions and have nothing to add here, write "See above."</i>
Recruitment	<i>Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.</i>
Ethics oversight	<i>Identify the organization(s) that approved the study protocol.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>

## Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

## ChIP-seq

## Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

## Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

## Files in database submission

Provide a list of all files available in the database submission.

## Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

## Methodology

## Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

## Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

## Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

## Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

## Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

## Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

## Flow Cytometry

## Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

## Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

## Instrument

Identify the instrument used for data collection, specifying make and model number.

## Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

## Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

## Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# Magnetic resonance imaging

## Experimental design

Design type	<i>Indicate task or resting state; event-related or block design.</i>
Design specifications	<i>Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.</i>
Behavioral performance measures	<i>State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).</i>

## Acquisition

Imaging type(s)	<i>Specify: functional, structural, diffusion, perfusion.</i>
Field strength	<i>Specify in Tesla</i>
Sequence & imaging parameters	<i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i>
Area of acquisition	<i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>
Diffusion MRI	<input type="checkbox"/> Used <input type="checkbox"/> Not used

## Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

## Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	<input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both
Statistic type for inference (See <a href="#">Eklund et al. 2016</a> )	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

## Models & analysis

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity
<input type="checkbox"/>	<input type="checkbox"/> Graph analysis
<input type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>





# Why fossil fuel producer subsidies matter

<https://doi.org/10.1038/s41586-019-1920-x>

Received: 20 November 2018

Accepted: 22 October 2019

Published online: 5 February 2020

 Peter Erickson<sup>1\*</sup>, Harro van Asselt<sup>2</sup>, Doug Koplow<sup>3</sup>, Michael Lazarus<sup>1</sup>, Peter Newell<sup>4</sup>,  
 Naomi Oreskes<sup>5</sup> & Geoffrey Supran<sup>5</sup>

 Arising from: J. Jewell et al. *Nature* <https://doi.org/10.1038/nature25467> (2019)

Around the globe, governments have pledged to remove support for coal, oil and gas, noting that such fossil fuel subsidies “undermine efforts to deal with climate change” by keeping greenhouse gas emissions higher than they otherwise would be<sup>1</sup>. Jewell et al.<sup>2</sup> used results of integrated assessment models to infer that eliminating subsidies would yield “limited emission reductions...except in energy-exporting regions”, and described the emission reduction benefits as “small”. This characterization is potentially misleading, and here we use a simple, sector-specific model to show how the emission reductions from producer subsidy reform could be more material than Jewell et al. suggest<sup>3</sup>. Fossil fuel producer subsidies delay a low-carbon transition in ways both material and political, and they deserve greater attention and transparency in global modelling analyses, as well as in policy-making.

The study by Jewell et al.<sup>2</sup> provides important findings related to fossil fuel subsidy removal. Using a synthesis of five Integrated Assessment Models (IAMs), they find that subsidy removal could reduce global emissions by 0.5 to 2 gigatonnes (Gt) of carbon dioxide (CO<sub>2</sub>) by 2030<sup>2</sup>. Jewell et al. characterize these global emission reductions as “unexpectedly small”, while noting that they would largely occur within a few energy-exporting countries and regions (Russia, the Middle East and Latin America)<sup>2</sup>.

We argue that the emissions reductions from subsidy removal are not small. By contrast, 0.5–2 Gt CO<sub>2</sub> amounts to roughly one quarter of the energy-related emission reductions pledged by all countries under the Paris Agreement (4–8 Gt CO<sub>2</sub>), all from a single policy approach that also comes with strong fiscal and other environmental benefits<sup>4</sup>. This scale of emission reductions should not necessarily be surprising or unexpected: few policy analysts hope that any single instrument can deliver reductions at the scale needed to meet climate goals.

Moreover, we argue that the impact of subsidy removal on emissions is likely to be more substantial than Jewell et al. find<sup>2</sup>, particularly when considering support for fossil fuel producers in high-income countries. Although their approach uses common IAM techniques, it does not adequately capture investment dynamics in the supply of new fossil fuels, and therefore misses a major pathway for subsidy reform to affect CO<sub>2</sub> emissions. Specifically, their approach does not consider how the timing of producer subsidies (concentrated early in an investment lifetime) and the higher effective discount rates of investors (as compared with society) affect investment decisions to bring on new supplies of oil.

Oil provides more of the world’s energy than any other fuel, and exploration and development of supplies remain robust<sup>5</sup>. The model in ref. <sup>2</sup> of producer subsidies to oil distributes regional subsidy totals equally to all oil fields—both new and already-producing fields—in each region, proportionate to annual output. However, that is often not how subsidies to oil producers work. Instead, governments frequently target subsidies more towards new capital investment than ongoing production. By lowering upfront cash flow requirements, government subsidies boost project investment metrics (such as rate of return or net present value), which leads producers to drill more new wells than they

would otherwise. This locks in higher future fossil fuel production and thus also higher future consumption and greenhouse gas emissions<sup>6</sup>.

Using the example of one type of subsidy for investment—accelerated depreciation of new capital investment—we illustrate how oil subsidies could have a bigger effect on global CO<sub>2</sub> emissions than in Jewell et al.’s analysis<sup>2</sup>. This particular form of support, exemplified by the intangible drilling cost (IDC) subsidy in the United States, allows companies to quickly write down capital investments that would otherwise depreciate more gradually, providing a boost to cash flow at the beginning of a project.

The IDC subsidy is underappreciated in Jewell et al.’s analysis<sup>2</sup> because they value it only at the reported value of about US\$0.20 per barrel (all dollar prices herein refer to 2016)<sup>7</sup>. This reflects the reduction in cash flow to the United States Treasury that results from the delay in annual tax payments. But whereas the USA government may be almost indifferent whether it receives tax revenues this year or the next, oil company investors are not, because they can use that cash flow to accelerate new investment.

If the IDC were valued not on a nominal cash basis but instead on a present value basis, using investor discount rates of 10% to 20%, the subsidy would make it substantially easier to invest in new oil fields, decreasing the breakeven oil price of new projects by US\$4 to \$7 per barrel (Table 1).

Changes in breakeven economics of this scale could have a substantial effect on global oil market price dynamics and consumption. This would especially be the case if subsidy removal were to render uneconomic many of the new projects on course to be developed before 2030. This outcome could well arise, since the USA has a substantial fraction (more than 40%) of the new oil projects that can be produced by 2030 (Extended Data Fig. 1). Other producers with substantial new supplies planned, such as Canada<sup>8</sup> and Norway<sup>9</sup>, also offer accelerated depreciation of new oil capital investments.

Table 1 estimates how the global oil market may respond to removal of the accelerated depreciation subsidies, based on a simple oil market model (see Methods). As shown, in the low-oil-price world featured by Jewell et al.<sup>2</sup>, the effect of removing the depreciation subsidy to producers could reduce global oil consumption by 440 to 770 million barrels in 2030.

Yet the previous analysis by Jewell et al.<sup>2</sup> includes only a very small fraction of this effect. They do not report this result, but we estimate it to be roughly 21 million barrels (Table 1, column A).

We therefore believe that, in their low oil price case, Jewell et al. missed a reduction in global CO<sub>2</sub> emissions from oil combustion on the order of 200 to 300 million tons CO<sub>2</sub> that could result from the removal of a single type of subsidy common in the USA and other oil-producing countries.

The actual outcome on net global CO<sub>2</sub> emissions from all fuels is likely to be somewhat lower because coal or gas might substitute for some

<sup>1</sup>Stockholm Environment Institute US, Seattle, WA, USA. <sup>2</sup>University of Eastern Finland, Law School, Joensuu, Finland. <sup>3</sup>Earth Track, Inc, Cambridge, MA, USA. <sup>4</sup>Centre for Global Political Economy, University of Sussex, Brighton, UK. <sup>5</sup>Department of the History of Science, Harvard University, Cambridge, MA, USA. \*e-mail: peter.erickson@sei.org

# Matters arising

**Table 1 | Removing subsidies that accelerate write-down of capital investment reduces global oil consumption**

	(A) Subsidy valued on cash basis, as in United States government source used by OECD and ref. <sup>2</sup>	Subsidy valued on present value basis at given investor discount rates		
		(B) Rate common in academic literature	(C) Rate common in industry studies	(D) Higher-risk rate (if weakened investor climate or higher-risk fields) <sup>10</sup>
		10%	15%	20%
<b>Effect of subsidy on economics of new oil projects</b>				
Effect on projects' breakeven price (US\$ per barrel)	0.20	4.20	5.80	7.30
<b>Market effects of subsidy removal for high-oil-price case in 2030</b>				
Increase in global oil price (US\$ per barrel)	0.07	1.40	1.90	2.40
Decrease in global oil consumption (millions of barrels)	4	76	110	130
Decrease in global CO <sub>2</sub> emissions from oil (millions of tonnes of CO <sub>2</sub> )	1	30	42	52
<b>Market effects of subsidy removal for low-oil-price case in 2030</b>				
Increase in global oil price (US\$ per barrel)	0.13	2.80	3.90	4.90
Decrease in global oil consumption (millions of barrels)	21	440	620	770
Decrease in global CO <sub>2</sub> emissions from oil (millions of tonnes of CO <sub>2</sub> )	8	180	250	310

These estimates of the effect of subsidies on projects' breakeven prices (first row) are calculated on a present value basis, as the production-weighted averages across nearly 800 discovered oil fields in the United States (see Methods). By contrast, Jewell et al.<sup>2</sup> value the fast depreciation subsidy only on a cash basis, spread across all fields; while they do not report the value of this subsidy in their analysis, we estimate it from the same primary sources they used to be about US\$ 0.20 per barrel (column A), as described further in the Methods. We estimate the market effects of removing these subsidies using a simple oil market model, at three different investor discount rates (columns B to D), all of which are on a nominal basis (no deduction for inflation). We assume that not-yet-developed USA oil projects are higher up the oil cost curve in 2030 (as is oil from other countries that also have a corresponding accelerated depreciation subsidy, like Canada or Norway), such that increases in the breakeven prices of these fields could well have a direct effect on long-term prices and consumption levels. We also assume here that subsidy removal begins immediately (in 2019), whereas Jewell et al. assume subsidy phase-out starts in 2020 and is completed in 2030. However, producer subsidy removal is not subject to the same concerns as consumer subsidy removal—namely equity and locked-in consumer behaviour—and thus would not need to be phased in so gradually.

of the lost oil consumption, though concurrent removal of subsidies for these fuels would minimize this effect. IAM models, like those used by Jewell et al.<sup>2</sup>, are well suited to evaluating these interactions. Yet the scale on which CO<sub>2</sub> emissions from oil have potentially been underestimated—equivalent to 10% to 60% of the reported global effect<sup>2</sup> due to removal of all subsidies (0.5 to 2 Gt CO<sub>2</sub> in 2030)—suggests that oil producer subsidies deserve greater attention and transparency in global modelling analyses.

The investment-oriented approach to modelling subsidies used here and the broader, average cost-curve approach of ref. <sup>2</sup> are not incompatible. Fossil fuel supply in IAMs could be modelled using an investment approach and vintage capital structure, as is often applied to power plants that have upfront costs and default lifetimes<sup>10</sup>. In such an approach, new oil deposits would also be modelled as prospective investments, as demonstrated here, using realistic discount rates of 10% to 20% that are common in the oil industry<sup>11</sup>.

In fact, subsidies may have an even more important role than we can quantify here. Extra company revenue resulting from subsidies can be used not only for more drilling, but also for product promotion, political activities and other efforts that fortify the industry's incumbent status. Subsidies also have a symbolic effect, in that they communicate the normative position that this industry and its activities are beneficial for society as a whole and, therefore, should be encouraged. Jewell et al.<sup>2</sup> disregard these socio-political effects when downplaying the value of removing fossil fuel subsidies.

The economic, political and symbolic effects of subsidies reinforce each other<sup>12</sup>. For example, subsidies can beget more subsidies, with new, long-lived fossil fuel infrastructure in turn (1) requiring further subsidization down the line to continue operating<sup>13,14</sup>, and (2) yielding beneficiaries who will vigorously defend continued subsidization<sup>15</sup>. Since there can be a revolving door between government staff and subsidy recipients, public officials may find it even harder to pass strong climate and energy policies<sup>16</sup>. Indeed, the most troubling impact and legacy of

fossil fuel subsidies may be the political barriers that fossil fuel producers have erected in recent decades against decarbonization efforts<sup>17,18</sup>.

Rapid low-carbon transitions consistent with the guardrails of the Paris Agreement require dramatically reduced fossil fuel production<sup>19</sup>. Subsidies to fossil fuel companies pose formidable financial, institutional and political obstacles to this transition, impeding the efficacy of greenhouse gas emission reduction strategies. The apparent small dollar values of producer subsidies in official, government-approved ledgers, and the limited emissions impact suggested by global models such as those used by Jewell et al.<sup>2</sup>, can be misleading. The actual impacts, particularly when one considers their social and political effects, are far greater.

Methods are in the Supplementary Information to this Matters Arising Comment.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The authors declare that data supporting the calculations in columns B through to D of Table 1 are included as Supplementary Information. The raw data analysed by the authors for Extended Data Fig. 1 are available from Rystad Energy in their UCube database, but restrictions apply to the availability of these data, which were used under license for the referenced study, and so are not publicly available. Raw data are available from the authors upon reasonable request and with permission of Rystad Energy.

## Code availability

No custom code or mathematical algorithms were used to generate results reported in this paper. The entirety of the oil market model is provided as equation (1) in the Methods.

1. Leaders of the G20 G20 Leaders' Statement: The Pittsburgh Summit. <https://www.oecd.org/g20/summits/pittsburgh/> (Organisation for Economic Co-operation and Development, 2009).
2. Jewell, J. et al. Limited emission reductions from fuel subsidy removal except in energy-exporting regions. *Nature* **554**, 229–233 (2018).
3. Erickson, P., Down, A., Lazarus, M. & Koplow, D. Effect of subsidies to fossil fuel companies on United States crude oil production. *Nat. Energy* **2**, 891–898 (2017).
4. Merrill, L., Gerasimchuk, I., Wooders, P. & Bassi, A. Fossil Fuel Subsidy Reform Research Suggests Emission Reductions Equivalent to at Least a Quarter of the Commitments Countries Made at Paris. <https://www.iisd.org/gsi/subsidy-watch-blog/fossil-fuel-subsidy-reform-research-suggests-emission-reductions-equivalent> (International Institute for Sustainable Development, 2018).
5. IEA World Energy Investment 2018 (Organisation for Economic Co-operation and Development, 2018).
6. Erickson, P. & Lazarus, M. Global emissions: new oil investments boost carbon lock-in. *Nature* **526**, 43 (2015).
7. OECD Companion to the Inventory of Support Measures for Fossil Fuels 2015 (Organisation for Economic Co-operation and Development, 2015).
8. Sawyer, D. & Stiebert, S. Fossil Fuels—At What Cost? Government support for upstream oil activities in three Canadian provinces: Alberta, Saskatchewan, and Newfoundland and Labrador <https://www.iisd.org/library/fossil-fuels-what-cost-government-support-upstream-oil-activities-three-canadian-provinces> (International Institute for Sustainable Development, 2010).
9. Erickson, P. & Down, A. How Tax Support For The Petroleum Industry Could Contradict Norway's Climate Goals <https://www.sei.org/publications/tax-petroleum-norways-climate-goals/> (Stockholm Environment Institute, 2017).
10. Iyer, G. C. et al. Improved representation of investment decisions in assessments of CO<sub>2</sub> mitigation. *Nat. Clim. Chang.* **5**, 436–440 (2015).
11. Fattouh, B., Poudineh, R. & West, R. Energy Transition, Uncertainty, and the Implications of Change in the Risk Preferences of Fossil Fuels Investors <https://www.oxfordenergy.org/publications/energy-transition-uncertainty-implications-change-risk-preferences-fossil-fuels-investors/?v=7516fd43adaa> (Oxford Institute for Energy Studies, 2019).
12. Seto, K. C. et al. Carbon lock-in: types, causes, and policy implications. *Annu. Rev. Environ. Resour.* **41**, 425–452 (2016).
13. Sovacool, B. K. Reviewing, reforming, and rethinking global energy subsidies: towards a political economy research agenda. *Ecol. Econ.* **135**, 150–163 (2017).
14. Newell, P. & Johnstone, P. The political economy of incumbency. In *The Politics of Fossil Fuel Subsidies and their Reform* (eds van Asselt, H. & Skovgaard, J.) 66–80 (Cambridge Univ. Press, 2018).
15. Koplow, D. Global energy subsidies: scale, opportunity costs, and barriers to reform. In *Energy Poverty* (eds Half, A., Sovacool, B. K. & Rozhon, J.) 316–337 (Oxford Univ. Press, 2014).
16. Oreskes, N. & Conway, E. M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming* (Bloomsbury Press, 2010).
17. Geels, F. W., Tyfield, D. & Urry, J. Regime resistance against low-carbon transitions: introducing politics and power into the multi-level perspective. *Theory Cult. Soc.* **31**, 21–40 (2014).
18. Supran, G. & Oreskes, N. Assessing ExxonMobil's climate change communications (1977–2014). *Environ. Res. Lett.* **12**, 084019 (2017).
19. Rogelj, J. et al. Mitigation pathways compatible with 1.5 °C in the context of sustainable development. In *Global Warming Of 1.5 °C: An IPCC Special Report On The Impacts Of Global Warming Of 1.5 °C Above Pre-Industrial Levels And Related Global Greenhouse Gas Emission Pathways, In The Context Of Strengthening The Global Response To The Threat Of Climate Change, Sustainable Development, And Efforts To Eradicate Poverty* Ch. 2 (IPCC, 2018).
20. Erickson, P. Confronting carbon lock-in: Canada's oil sands. SEI discussion brief. <https://www.sei.org/publications/confronting-carbon-lock-canadas-oil-sands/> (Stockholm Environment Institute, 2018).

**Acknowledgements** P.E. and M.L. thank A. Vogt-Schilb, S. Pye and N. Bauer for discussions about IAM models. P.E. acknowledges funding from the Schmidt Family Foundation.

**Author contributions** P.E. and M.L. conceptualized the research (with input from H.v.A., D.K., N.O. and G.S.). P.E. carried out the numerical modelling. P.E. wrote and revised the manuscript (with contributions from H.v.A., D.K., M.L., P.N., N.O. and G.S.).

**Competing interests** The authors declare no competing interests.

#### Additional information

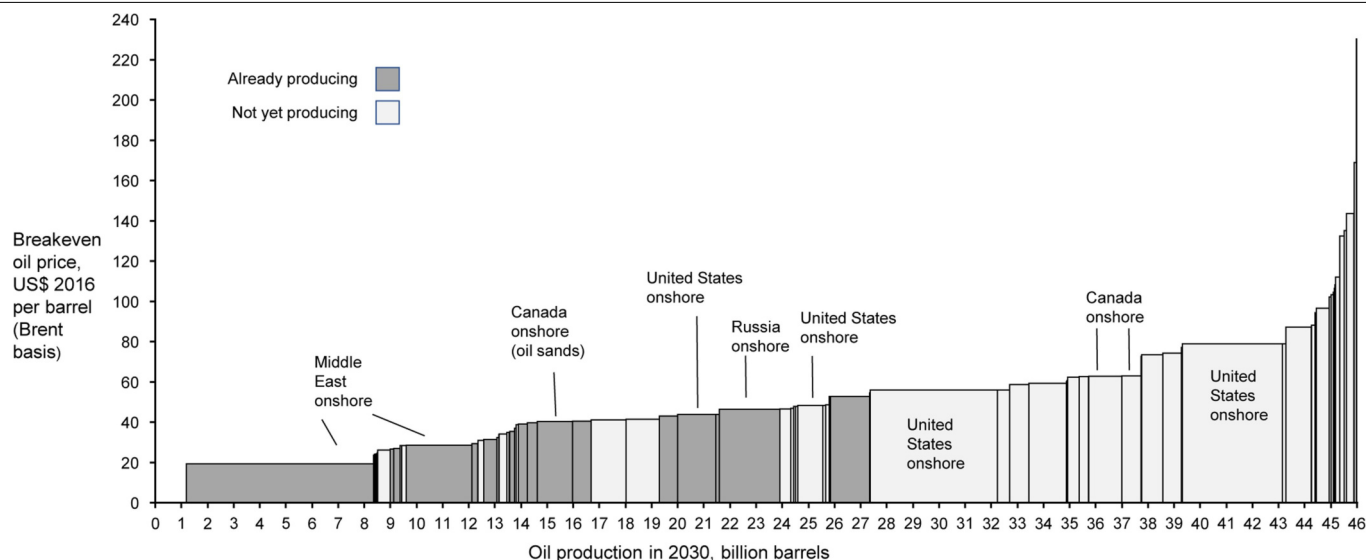
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1920-x>.

**Correspondence and requests for materials** should be addressed to P.E.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020



**Extended Data Fig. 1 | Cost curve of world oil production in 2030.** The cumulative supply of oil in 2030 is shown for increasing oil price. Most blocks (64 of 80) in this cost curve represent a combination of a particular stage of development (one of four) in eight major world regions (the continents plus the Middle East and Russia minus Antarctica), whether onshore or offshore.

Further (16) blocks represent the USA or Canada, since they are major new sources of oil (about 41% and 7% of all regions that are not yet producing oil). The figure is adapted from figure 1 in ref.<sup>20</sup> and based on data from Rystad Energy (see 'Data availability' section).

# Reply to: Why fossil fuel producer subsidies matter

<https://doi.org/10.1038/s41586-019-1921-9>

Published online: 5 February 2020

Jessica Jewell<sup>1,2,3,4\*</sup>, Johannes Emmerling<sup>5,6</sup>, Vadim Vinichenko<sup>2,3</sup>, Christoph Bertram<sup>7</sup>, Loïc Berger<sup>5,6,8</sup>, Hannah E. Daly<sup>9</sup>, Ilkka Keppo<sup>10</sup>, Volker Krey<sup>11,12</sup>, David E. H. J. Gernaat<sup>13,14</sup>, Kostas Fragkiadakis<sup>15</sup>, David McCollum<sup>16</sup>, Leonidas Paroussas<sup>15</sup>, Keywan Riahi<sup>11,17</sup>, Massimo Tavoni<sup>5,6,18</sup> & Detlef van Vuuren<sup>13,14</sup>

Replying to: P. Erickson et al. *Nature* <https://doi.org/10.1038/s41586-019-1920-x> (2020)

In 2009, the G20 countries pledged to phase out fossil fuel subsidies<sup>1</sup>. Our original Letter highlighted that about 95% of subsidies go to consumers and two-thirds are in the Middle East, Russia and Latin America<sup>2</sup>. We also found the largest emission reductions from subsidy removal would occur in those three regions, where low oil prices provided a unique political opportunity and the reforms would harm fewer poor people. In the accompanying Comment<sup>3</sup>, Erickson et al. argue that we downplay the impact of subsidy removal and the effect of subsidies for oil producers, such as the USA's intangible drilling cost (IDC) scheme. Here we show large variations in such schemes and estimate their impact to be within the range of the sensitivity analysis from our original article. The USA IDC may represent a unique political opportunity for producer subsidy reform, but reforming such schemes may be counterproductive in countries where they are applied in tandem with high taxes for oil production.

We estimated that emission reductions from subsidy removal would be between 2–8% and 3–15% of those required by 2030 to achieve the 1.5 °C and 2 °C targets. We called this “unexpectedly small” because these estimates contrast with sweeping statements that subsidy removal would have “significant”<sup>4</sup> effects and is “the missing link in the fight against climate change”<sup>5</sup>. Yet we agree with Erickson et al.<sup>3</sup> that given the immensity of the climate challenge, these numbers are notable and are certainly not an argument against subsidy reform.

Erickson et al.<sup>3</sup> estimate the size and effect of the USA's IDC scheme, which allows accelerated depreciation of drilling costs, essentially tax deferrals for oil producers. Their approach is different from ours in how subsidies are defined and measured. In our original Letter<sup>2</sup>, we used government inventories<sup>6–8</sup> for our central estimate, because these are the very subsidies that governments have pledged to remove. Erickson et al.<sup>3</sup> consider any regulation that makes oil production more profitable to be a subsidy even if it does not involve net transfers from the government. This leads them to use data not from government inventories of subsidies but from analysing oil production economics. Thus, Erickson et al.<sup>3</sup> analyse the hypothetical cash flow for 800 oil fields in the USA and calculate the effect of the IDC scheme on the breakeven price of individual projects—we call this the ‘effective subsidy rate’. They then assess the global impact of similar schemes by assuming

all oil producers worldwide benefit from the same effective subsidy rate as in the USA.

Global IAMs can greatly benefit from such data if they are parameterized for long-term global scenarios. The first set of parameters defines how accelerated depreciation affects the effective subsidy rate. This depends on a project's breakeven price, discount rate, share of capital costs, the national tax regime, and the design of the accelerated depreciation scheme, all of which vary widely across countries and over time (Methods). To determine whether Erickson et al.'s results<sup>3</sup> would affect our original findings<sup>2</sup>, we developed a discounted cash flow model to analyse the effective subsidy rate for the USA IDC and from accelerated depreciation schemes for three additional countries with diverse institutional arrangements and geographies (Methods).

In the case of the USA, our model provides results similar to those of Erickson et al.<sup>3</sup> for the 2016 case, but the 2017 tax cut reduced the effective subsidy rate by about half and the recent fall in the cost of North American tight oil reduced it by another 30% (Table 1, Methods). The effective subsidy rates from accelerated depreciation schemes in Canada, Norway and Russia under a range of plausible breakeven prices are between two and ten times smaller than the USA 2016 case. Using this range, we estimate the global effective subsidy rate from accelerated depreciation schemes to be US\$0.3–1.9 per barrel [using central assumptions, as described in the Methods, the value is US\$1.0; central values are shown herein in square brackets] (US dollar prices herein refer to 2016; Table 1).

The uncertainty in estimating production subsidies is well known<sup>6,9,10</sup>. That is why, in our original Letter<sup>2</sup>, we included a sensitivity analysis using an alternative estimate of production subsidies (including an alternative calculation of the USA's IDC scheme)<sup>10</sup>. Oil production subsidies in that analysis for the low-oil-price scenario were about fifteen times higher than those reported in government inventories<sup>2</sup>. The oil production subsidy rates in that original sensitivity analysis are also generally higher than the effective subsidy rates we estimate for accelerated depreciation schemes (column C of Table 1).

The second step in the analysis by Erickson et al.<sup>3</sup> is to estimate the effect of accelerated depreciation schemes on global oil consumption

<sup>1</sup>Division of Physical Resource Theory, Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden. <sup>2</sup>Centre for Climate and Energy Transformations, University of Bergen, Bergen, Norway. <sup>3</sup>Department of Geography, Faculty of Social Sciences, University of Bergen, Bergen, Norway. <sup>4</sup>Risk and Resilience Program, International Institute for Applied Systems Analysis, Laxenburg, Austria. <sup>5</sup>RFF-CMCC European Institute on Economics and the Environment, Milan, Italy. <sup>6</sup>Fondazione Centro Euromediterraneo sui Cambiamenti Climatici, Lecce, Italy. <sup>7</sup>Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany. <sup>8</sup>IESEG School of Management, CNRS, Université Lille, UMR 9221-LEM, Lille, France. <sup>9</sup>MaREI, the SFI Research Centre for Energy, Climate and Marine, Environmental Research Institute, University College Cork, Cork, Ireland. <sup>10</sup>UCL Energy Institute, University College London, London, UK. <sup>11</sup>Energy Program, International Institute for Applied Systems Analysis, Laxenburg, Austria. <sup>12</sup>Industrial Ecology and Energy Transitions Programmes, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. <sup>13</sup>Copernicus Institute for Sustainable Development, University of Utrecht, Utrecht, The Netherlands. <sup>14</sup>PBL Netherlands Environmental Assessment Agency, The Hague, The Netherlands. <sup>15</sup>National Technical University of Athens, Athens, Greece. <sup>16</sup>Electric Power Research Institute, Palo Alto, CA, USA. <sup>17</sup>Institute of Thermal Engineering, Graz University of Technology, Graz, Austria. <sup>18</sup>Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan, Italy.

\*e-mail: [jewell@chalmers.se](mailto:jewell@chalmers.se)



# Matters arising

**Table 1 | The effect of oil production subsidies on producer costs and global oil consumption**

	IAM analysis of all producer subsidies in Jewell et al. <sup>2</sup> (low-oil-price scenario)		Discounted cash flow model of accelerated depreciation schemes (10% discount rate except for final row; see Methods)	
	(A) Main estimate of production subsidies from ref. <sup>6</sup>	(B) Higher production subsidies from refs. <sup>9,10</sup>	(C) Variations in tax rates, capital cost, accelerated depreciation schemes, breakeven prices and elasticities (our analysis)	(D) Erickson et al. <sup>3</sup>
<b>Effective production subsidy rate (US\$ per barrel)</b>				
USA	0.6	2.4	1.9 (2019 case) 4.9 (2016 case)	4.2 (2016 case)
Other regions	Canada 1.1 Europe 0.4 Russia 0 MENA 0	Canada 1.5 Europe 2.2 Russia 5.2 MENA 2.3	Canada 0.5–1.4 [0.9] Norway 0.9–2.0 [1.5] Russia 0.9–2.1 [1.6] Saudi Arabia and Nigeria 0	
Global	0.2	2.6	0.3–1.9 [1.0]	
<b>Change in global oil extraction or consumption (millions of barrels per year) for the low-oil-price scenario</b>				
Change in extraction due to higher production subsidy estimate		590		440
Variation due to effective subsidy rate using elasticities in Erickson et al. <sup>3</sup> and 10% discount rate			30–200 [110]	
Variation due to elasticity assumptions <sup>11,12</sup> using central effective subsidy rate and 10% discount rate			20–140 [90]	
Variation due to discount rates using central effective subsidy rate and elasticities in Erickson et al. <sup>3</sup> In column (C), discount rates vary from 7.5% to 20% and in column (D) from 10% to 20% (Methods). In both cases, the discount rate for the central estimate is 15%.			90–150 [130]	440–770 [620]
Columns A and B contain estimates of all oil production subsidies from our original article <sup>2</sup> . In our sensitivity case, the subsidy rate and its effect on oil production is higher than under accelerated depreciation schemes (column C). For the USA, the results for the 2016 case are in italics including from Erickson et al. (column D). ‘Canada’ refers to the CAJAZ (Canada, Japan, Australia and New Zealand) region whose oil production is dominated by Canada (over 98%). ‘MENA’ refers to the Middle East and North Africa region.				

with a simple oil market model. Their calculation is sensitive to supply and demand elasticities that are highly uncertain (Methods). They use a single value for demand elasticity and a single value for supply elasticity for each oil price. A range of supply and demand elasticities from previous studies that used the same simple oil market model<sup>11,12</sup> changes the results by almost by an order of magnitude even under the same effective subsidy rate (Table 1, Methods).

In the sensitivity analysis from our original article, we estimated that a more than tenfold increase in oil production subsidies would increase oil extraction by 590 million barrels per year (Table 1). The higher production subsidies (including all production subsidies, not just oil) would increase emission reductions from subsidy removal by 0.3 gigatonnes of carbon dioxide per year in 2030, which is about 13% higher than the main estimate of the model used for that sensitivity analysis, or about 1% of the emission reduction required by 2030 to achieve the 1.5 °C or 2 °C target (Methods).

The final parameter affecting the effective subsidy rate is the discount rate, which Erickson et al.<sup>3</sup> assume varies between 10–20%. The upper end of this range is speculative because discount rates for the oil sector have generally varied between 9% and 11%<sup>13</sup> (Methods). Table 1 shows our results using a discount rate of 10%, but our conclusions are robust over the full range in Erickson et al.<sup>3</sup>: a 20% discount rate increases the global effective subsidy rate to US\$0.4–2.7 [1.4] per barrel (Methods).

This exchange highlights the importance of improving IAM parameters by incorporating new data. Such data are more meaningful to global long-term IAMs if it is clear whether and how they are applicable beyond a single country at a single point in time. The generalizability of such data can be improved if they extend to a wider and more representative sample<sup>9,10</sup>, which IAMs can use, as illustrated by the sensitivity analysis in our original article. Finally, these data should be up-to-date and transparent about uncertainties, including those arising from differences in policy environments.

Although the effect of accelerated depreciation schemes can be incorporated into IAMs by adjusting the effective subsidy rate, we also agree with Erickson et al.<sup>3</sup> that IAMs should better represent oil and gas infrastructure in the same way as they model the vintage structure of

the power sector<sup>14</sup>. Another promising avenue would be to depict oil and gas investments using a real options ‘wait and see’ approach<sup>15</sup> and to model price formation in the oil market<sup>16</sup> more realistically. These improvements may either dampen or amplify the effects of subsidies in IAMs, depending on whether infrastructural inertia, ‘wait and see’ behaviour, and strategic markets are more or less responsive to producer cost signals than in today’s IAMs.

We also strongly agree with Erickson et al.<sup>3</sup> that the social and political impacts of subsidy removal should always be examined in tandem with their emission impacts. However, it is time for social scientists to go beyond listing various negative effects of subsidies which are well documented in the literature and clearly extend beyond economics<sup>17–20</sup> and instead identify opportunities and pathways for reform. That is why, in our original article, we complemented energy and emissions analysis with a discussion of the socio-political impacts of subsidies to identify a political opportunity for reform in oil- and gas-exporting countries under low oil prices where reducing consumption subsidies would affect fewer poor people, relieve squeezed government budgets and lead to the largest emission reductions.

A lesson from our original article is that the environmental and socio-political impacts of and obstacles to consumer subsidy reform vary between countries. This is almost certainly the case for producer subsidies as well. In the USA, the original rationale (energy security and uncertainty in oil drilling) for the IDC is outdated, and the scheme now does little more than confer an unfair advantage on a polluting, privately owned and profitable industry. Reforming this scheme is complicated by the political clout of the industry, but at least its public benefits and endpoint are clear.

However, such subsidies are much more difficult to identify, much less reform, in countries like Norway and Russia where oil producers pay very high taxes—reaching over 70% on profits. These taxes are a major source of government revenue used to fund public services. Would reforming accelerated depreciation schemes in these contexts also mean tax reductions for the industry? Would the endpoint be to bring the oil industry in line with the rest of the economy, something clearly not desirable either socially or environmentally? And if not, what would be the goal and the strategy for reform?

Generalizing insights from the USA to the whole world is misleading both in terms of science and policy. Finding effective strategies to meet the Paris Agreement requires a detailed understanding of how oil production and other carbon-intensive sectors are embedded in national socio-political and economic contexts.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The authors declare that the data supporting the calculations are available in the Methods or from publicly available sources cited in the Methods.

## Code availability

No custom code or algorithms were developed for the discounted cash flow results reported in this paper.

1. Joint Report By IEA, OPEC, OECD And World Bank On Fossil-fuel And Other Energy Subsidies: An Update Of The G20 Pittsburgh And Toronto Commitments <https://www.oecd.org/env/49090716.pdf> (IEA, OPEC, OECD and World Bank, 2011).
2. Jewell, J. et al. Limited emission reductions from fuel subsidy removal except in energy-exporting regions. *Nature* **554**, 229–233 (2018).
3. Erickson, P. et al. Why fossil fuel producer subsidies matter. *Nature* <https://doi.org/10.1038/s41586-019-1920-x> (2020).
4. Intergovernmental Panel on Climate Change (IPCC). Technical Summary. In *Climate Change 2014: Mitigation of climate change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Edenhofer, O. et al.) 33–107 (Cambridge Univ. Press, 2014).
5. Friends of Fossil Fuel Subsidy Reform. *Briefing Note July 2015: Fossil Fuel Subsidy Reform and the Communiqué*. <http://ffsr.org/wp-content/uploads/2015/07/ffsr-communication-briefing-note.pdf> (Friends of Fossil Fuel Subsidy Reform, 2015).
6. OECD *OECD Companion to the Inventory of Support Measures for Fossil Fuels 2015*. <https://www.oecd.org/publications/oecd-companion-to-the-inventory-of-support-measures-for-fossil-fuels-2015-9789264239616-en.htm> (OECD, 2015).
7. IEA *World Energy Outlook 2016* <https://webstore.iea.org/world-energy-outlook-2016> (International Energy Agency, 2016).
8. IEA *World Energy Outlook 2014* <https://webstore.iea.org/world-energy-outlook-2014> (International Energy Agency, 2014).
9. Bast, E., Doukas, A., Pickard, S., Burg, L. Van Der & Whitley, S. *Empty Promises: G20 Subsidies to Oil, Gas And Coal Production*. <https://www.oecd.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9957.pdf> (Overseas Development Institute and Oil Change International, 2015).

10. Gerasimchuk, I. et al. Zombie energy: climate benefits of ending subsidies to fossil fuel production. Working paper. <https://www.iisd.org/sites/default/files/publications/zombie-energy-climate-benefits-ending-subsidies-fossil-fuel-production.pdf> (International Institute for Sustainable Development, Global Subsidies Initiative, and Overseas Development Institute, 2017).
11. Erickson, P., Down, A., Lazarus, M. & Koplow, D. Effect of subsidies to fossil fuel companies on United States crude oil production. *Nat. Energy* **2**, 891–898 (2017).
12. Erickson, P. & Lazarus, M. Impact of the Keystone XL pipeline on global oil markets and greenhouse gas emissions. *Nat. Clim. Chang.* **4**, 778–781 (2014).
13. Damodaran, A. *Cost Of Equity And Capital (Updateable)* [http://people.stern.nyu.edu/adamodar/New\\_Home\\_Page/datafile/wacc.htm](http://people.stern.nyu.edu/adamodar/New_Home_Page/datafile/wacc.htm) (2019).
14. Johnson, N. et al. Stranded on a low-carbon planet: implications of climate policy for the phase-out of coal-based power plants. *Technol. Forecast. Soc. Change* **90**, 89–102 (2015).
15. Compennolle, T., Welkenhuysen, K., Huismans, K., Piessens, K. & Kort, P. Off-shore enhanced oil recovery in the North Sea: the impact of price uncertainty on the investment decisions. *Energy Policy* **101**, 123–137 (2017).
16. Ansari, E. & Kaufmann, R. K. The effect of oil and gas price and price volatility on rig activity in tight formations and OPEC strategy. *Nat. Energy* **4**, 321–328 (2019).
17. Victor, D. G. *The Politics of Fossil-Fuel Subsidies* [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1520984](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1520984) (Global Subsidies Initiative and International Institute for Sustainable Development, 2009).
18. Inchauste, G. & Victor, D. G. *The Political Economy of Energy Subsidy Reform Public Sector Governance* (World Bank, 2017).
19. Sovacool, B. K. Reviewing, reforming, and rethinking global energy subsidies: towards a political economy research agenda. *Ecol. Econ.* **135**, 150–163 (2017).
20. Lockwood, M. Fossil fuel subsidy reform, rent management and political fragmentation in developing countries. *New Polit. Econ.* **20**, 475–494 (2015).

**Acknowledgements** This work was supported by the Research Council Norway under the Contractions project (“Analyzing past and future energy industry contractions: towards a better understanding of the flip-side of energy transitions”) under grant agreement number 267528/E10. We thank A. Cherp for discussions on the discounted cash flow model.

**Author contributions** The change in composition (removal of Nawfal Saadi and addition of Hannah Daly) and order of the author list in this Matters Arising compared to the original Letter reflects the contributions to the Matters Arising Reply, which relied on a discounted cash flow model and additional empirical research in order to validate the assumptions and sensitivity analysis from the original article. J.J., J.E., V.V. and C.B. wrote the Reply with contributions from L.B., H.D., I.K., V.K., D.E.H.J.G., K.F., D.M., L.P., K.R., M.T. and D.V. The numerical model was conceived and designed by J.J. and V.V. and implemented by V.V. The results were analysed by J.J. and V.V.

**Competing interests** The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1921-9>.

**Correspondence and requests for materials** should be addressed to J.J.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

# Author Correction: H2A.Z facilitates licensing and activation of early replication origins

---

<https://doi.org/10.1038/s41586-020-1948-y>

---

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1877-9>

---

Published online 25 December 2019

---

Haizhen Long, Liwei Zhang, Mengjie Lv, Zengqi Wen, Wenhao Zhang, Xiulan Chen, Peitao Zhang, Tongqing Li, Luyuan Chang, Caiwei Jin, Guozhao Wu, Xi Wang, Fuquan Yang, Jianfeng Pei, Ping Chen, Raphael Margueron, Haiteng Deng, Mingzhao Zhu & Guohong Li

---

In this Article, ARSs should have been defined as ‘autonomously replicating sequences’, not ‘automatic replication sequences’. This error has been corrected online.

# Publisher Correction: Dietary salt promotes cognitive impairment through tau phosphorylation

---

<https://doi.org/10.1038/s41586-019-1925-5>

---

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1688-z>

---

Published online 23 October 2019

---

**Giuseppe Faraco, Karin Hochrainer, Steven G. Segarra,  
Samantha Schaeffer, Monica M. Santisteban, Ajay Menon,  
Hong Jiang, David M. Holtzman, Josef Anrather &  
Costantino Iadecola**

---

In Fig. 1d of this Article, owing to an error in the production process, an asterisk on the bracket for the 12-weeks (12w) time-point is missing. In addition, Fig. 3 contains errors present at submission. In Fig. 3f in the calpain activity panel, the two data points at  $0.00 \times 10^3$  relative fluorescence units (RFU) per mg for the normal diet (ND) group should not be there. In the two rightmost panels of Fig. 3h, both of the first two bars should be blue (indicating the wild type (WT)) and both of the second two bars should be red (indicating eNOS<sup>-/-</sup> or nNOS<sup>-/-</sup>, respectively). These errors have been corrected online.

---

# **Publisher Correction: Nanomagnetic encoding of shape-morphing micromachines**

---

<https://doi.org/10.1038/s41586-019-1888-6>

---

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1713-2>

---

Published online 6 November 2019

---

**Jizhai Cui, Tian-Yun Huang, Zhaochu Luo, Paolo Testa, Hongri Gu,  
Xiang-Zhong Chen, Bradley J. Nelson & Laura J. Heyderman**

---

In the HTML version (the PDF and print versions were correct) of this Article, owing to a typesetting error, Tian-Yun Huang should not have been affiliated with the Laboratory for Multiscale Materials Experiments, Paul Scherrer Institute, Villigen, Switzerland (affiliation 2). This error has been corrected online.



# **Publisher Correction: TGF- $\beta$ orchestrates fibrogenic and developmental EMTs via the RAS effector RREB1**

---

<https://doi.org/10.1038/s41586-020-1956-y>

---

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1897-5>

---

Published online 8 January 2020

---

Jie Su, Sophie M. Morgani, Charles J. David, Qiong Wang,  
Ekrem Emrah Er, Yun-Han Huang, Harihar Basnet, Yilong Zou,  
Weiping Shu, Rajesh K. Soni, Ronald C. Hendrickson,  
Anna-Katerina Hadjantonakis & Joan Massagué

---

In Fig. 4j of this Article, the label on the bottom panel should have been '*Rreb1*<sup>-/-</sup> ESC' instead of '*Rreb1*<sup>+/+</sup> ESC'. This error has been corrected online.



ADAPTED FROM GETTY

## OUT-OF-OFFICE REPLIES AND WHAT THEY SAY ABOUT YOU

An automated e-mail response posted on Twitter unleashed a social-media debate about the importance of work–life balance. **By Stephana Cherak**

**O**ne weekend last October, I received an out-of-office reply from an academic faculty member: “I do not respond to e-mails on weekends. If this is an emergency, please call my mobile. If you do not have my mobile number, then you do not have a weekend emergency.” The tone, I felt, was telling – researchers are often under pressure to keep up with immense workloads and to stay on top of the changing world of science, while trying to protect their

life outside work from being invaded by faux emergencies or apparently urgent e-mails.

Indeed, in response to *Nature*’s 2019 biennial

**“The struggle to balance work and life is a particular challenge for graduate students.”**

PhD survey, only 37% of PhD students agreed that their institute supported a good work–life balance. Thirty-four per cent felt it failed to do so, and nearly 40% of respondents said they were unsatisfied with their work–life balance.

Researchers, including PhD students, senior scientists and graduate students like myself, are not the only ones facing pressure to work harder and faster. One 2017 survey (see [go.nature.com/2rxh4rb](https://go.nature.com/2rxh4rb)) of workers in Britain, for example, found that nearly half said

## Tales from Twitter

**When I tweeted an out-of-office message in October that read: “I do not respond to e-mails on weekends,” the post went viral. So far, it has received 41,200 likes and more than 4,600 re-tweets — nearly 3.1 million Twitter accounts have interacted with the post. I’ve curated a selection of my favourite replies so far.**

### Lives aren’t controlled by e-mail

“I had a discussion with a busy academic recently about an international trip where she was unexpectedly unable to access e-mail and it was so backlogged by her return that she just ignored/ deleted all unopened mail. And then marvelled that the world really didn’t stop.” @JessicaRenee\_83

“When I took parental leave recently, I didn’t respond to e-mails for almost an entire month. The world didn’t end. I’d gladly do it again.” @tomkXY

### Honouring boundaries and timelines

“I haven’t quite reached this level of firmness in my boundaries, but I do know that my life has gotten much better since I decided that I don’t need ‘fastest/ best/ most consistent e-mail responder’ to be part of my professional legacy.” @popmediaprof

“I used one of these responders for years as a prof. Now I just take my sweet time to reply to e-mails and people learn. Their timelines aren’t my timelines.” @DoctorLindy

### Do as I say, not as I do

“Fully support this. I also recommend adding a statement to any e-mails sent out of regular business hours to the effect of ‘this email is being sent out of normal business hours and I don’t expect an immediate reply’.” @runforbooze

“On weekends I respond to e-mails on my own timeline. Sometimes that means right away, but most times it means not until Monday. If a student reads too much into a quick reply and comes to expect that, I simply say that past performance is no guarantee of future results.” @Meteodan

### And the winner is ...

“There is no such thing as an academic emergency.” @JLasaiane

their jobs require them to work very hard — compared with less than one-third in 1992 (see ‘Off balance’). When I tweeted the automated response in October, it went viral, suggesting work–life balance is an issue throughout the working world, not just in academia (see ‘Tales from Twitter’).

Forgoing work–life balance in order to be as productive as possible is tempting in academia, because those who are the most productive are also the most rewarded. Although hard work might be valued, it’s hardly the only factor in a successful career — numerous studies have shown that as we work more, the quality of what we do decreases. The scientific community should learn to value quality over quantity.

The struggle to balance work and life is a particular challenge for graduate students. They often find the line between the two becomes blurred with efforts to seem productive and important. Institutions have a responsibility to students and employees to protect their health; often this means helping people to manage their work–life balance. But the culture at science institutes can tend towards long working hours. As Meghan Duffy, an ecologist at the University of Michigan in Ann Arbor, told *Nature* in 2017: “The idea that you have to put in long hours is pervasive. If you’re not working 60 or 80 hours a week, you’re not doing enough. It makes people insecure.”

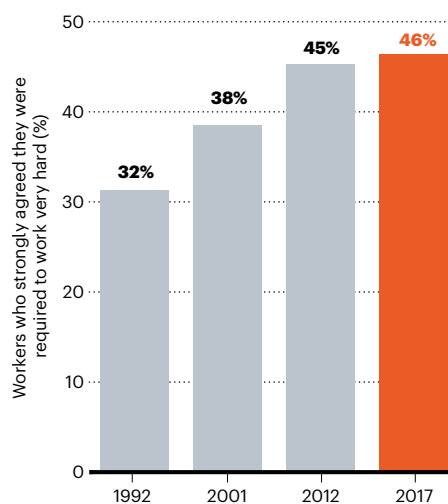
After receiving hundreds of comments in response to my tweet about the efforts people go to to maintain balance, I was prompted to reflect on my own techniques. Here are a few of the principles that I’ve adopted as I strive to balance work and other aspects of life:

**Don’t rush, set your own pace.** Graduate students often want immediate results. I thought the more courses I took or the more experiments I planned in one week, the earlier I might graduate. During my undergraduate degree I started arriving at university at 6 a.m. and leaving at 8 p.m. But, in graduate school, I soon learnt my studies were not a sprint, but a marathon. Make that marathon work for you. During my master’s degree I was also competing at a high level in athletics; this meant my course-work needed to be restructured so I could be successful at both. In every department there are resources to help you to personalize your programme. Take advantage of these opportunities, and don’t feel the need to align with the status quo — you can save yourself a lot of sleepless nights by planning strategically.

**Ask for support.** Academics often think the work they do should be perfect. This causes trouble: graduate students often neglect their lives outside work to strive for perfection. But as a student, you are expected to ask for help, assistance and guidance. Finding the confidence to ask for support from supervisors and mentors is beneficial and necessary for

## OFF BALANCE

Nearly half of UK workers across all sectors say their job requires them to work very hard.



SKILLS AND EMPLOYMENT SURVEY/CARDIFF UNIVERSITY/ UNIVERSITY COLLEGE LONDON/ UNIVERSITY OF OXFORD

growth. Academic studies can mean working alone, but it is important to remember that one of the most rewarding aspects of being a researcher is collaborating with others. Many would say that you are only as strong as those you surround yourself with.

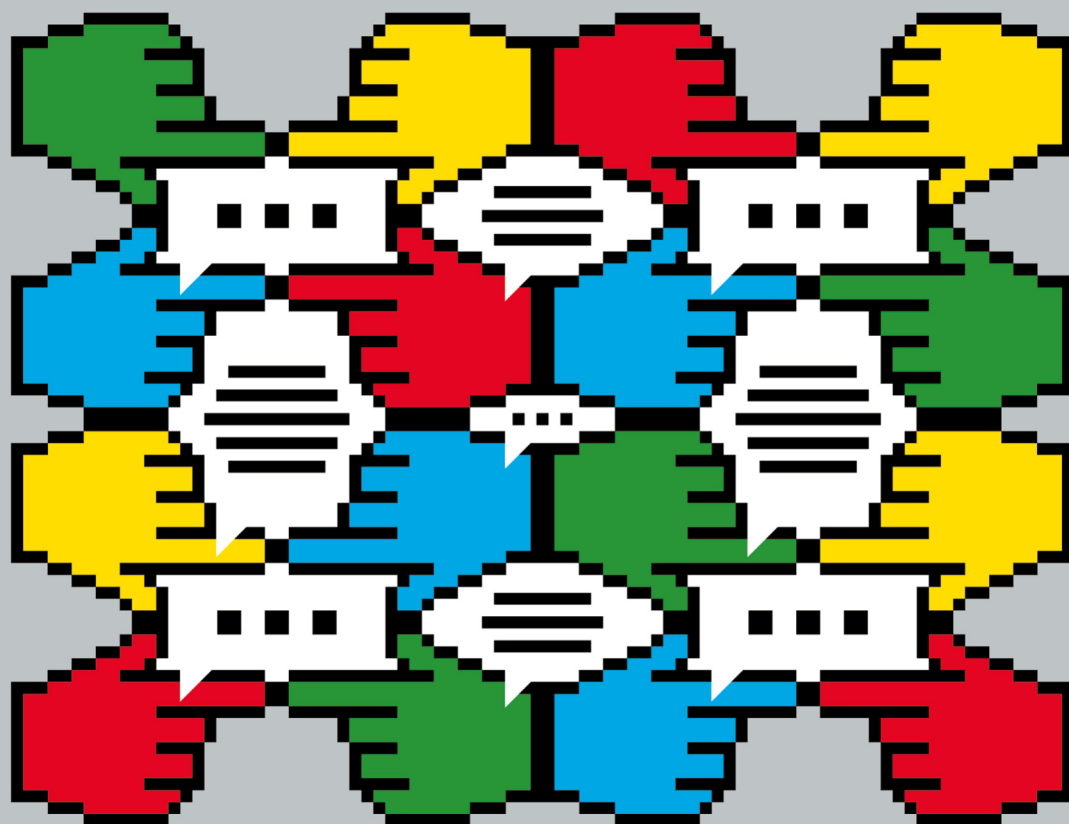
**Define your own personal balance.** Graduate students often spread themselves too thinly, agreeing to spend longer hours in the lab or in

**“I thought the more courses I took or the more experiments I planned, the earlier I might graduate.”**

front of a laptop at the expense of their time outside work. Although there are certain non-negotiable expectations and levels of professionalism required for a graduate degree, attaining these shouldn’t come at the cost of your health and well-being. Dedicate time to defining and finding your personal balance, and honour this throughout your career.

**Always be kind.** Academia has rough edges and the way that researchers communicate can often seem curt or unpleasant for someone who isn’t used to it. Seek to pursue your own acts of academic kindness. Perhaps thank colleagues more for their hard work — or even send a card. This might inspire others to shift their own behaviour and perhaps to be a bit more generous and compassionate. By being kind ourselves, we might help to make academia, as a whole, just a little bit kinder.

**Stephana Cherak** is a PhD student in the Department of Community Health Sciences, University of Calgary, Canada. Her PhD research focuses on patient and family-centred critical care.



# TECH TOOLS TO MAKE RESEARCH MORE OPEN AND INCLUSIVE

Laboratory heads are deploying apps and software in innovative ways to build broad and diverse research groups. **By Kendall Powell**

**D**oris Taylor knows the sting of being set apart as different. As a young, lesbian woman starting her career in regenerative-medicine research in the late 1980s, she was often excluded from faculty functions and private meetings on the golf course. “You want to be differentiated when doing great science, but not because of who you are,” she says.

Galvanized by her experiences, Taylor has built a laboratory group at the Texas Heart Institute in Houston that strives to be diverse and culturally sensitive. She knew she had come close when she overheard an undergraduate researcher telling his mother about a lab birthday celebration: “I was the only white guy there — it was great!”

Taylor thought carefully about how best to

build a diverse, inclusive and equitable team, representing a range of perspectives and backgrounds, and like many other investigators who value such things, she has increasingly relied on technology to advance those goals.

Group leaders say that these tools can help to flatten power differentials between lab members and keep people connected and communicating on common, and importantly, even ground. The tools are familiar, and even ubiquitous — Slack, Skype and WhatsApp (Taylor’s tool of choice), for example. But when deployed strategically, these apps can promote a more level playing field to benefit colleagues from disadvantaged and under-represented backgrounds, those with disabilities, or those who might work and think differently.

That’s not to say technology is a silver bullet — building an inclusive environment requires a sustained commitment from lab leaders and members, on multiple levels and using many techniques. And no amount of technology can erase bullying, discrimination and other bad behaviours from the workplace. But these tools are helping many inclusive-minded group leaders to transform research from an isolated pursuit into a more open, collective exercise.

“Any technology that increases communication in a way that is non-threatening, is beneficial,” says Taylor.

## Messaging equality

The University of Helsinki’s Computational Field Theory Group, which studies what happened in the 10 picoseconds after the Big Bang,



## Work / Technology & tools

for instance, uses an open-source messaging platform similar to Slack to share data and discuss results with collaborators. But the group takes the tool, called Mattermost, even further: it uses it as a forum for nearly all group communications, from discussing research projects to organizing spontaneous outings and lunches. This keeps discussions open and transparent to all of the group's 20 or so members and their colleagues. Members frequently add notes from face-to-face conversations as a transcript record and to keep everyone in the know.

David Weir, a physicist in the group, explains that workplace surveys had revealed internal communication problems in the Helsinki physics department, and showed that members, particularly women who didn't speak Finnish fluently, often felt isolated. "I do think [Mattermost] helps lower the threshold to people participating," he says.

Saga Säppi, a PhD student in a neighbouring theoretical-physics group, says the open-to-all messaging has made a "night and day" positive difference to social interactions. And it lowered the barrier to getting help, she adds, by making it easier to send research questions informally to the entire group rather than having a time-consuming e-mail exchange with a supervisor.

### Terms of engagement

Other tools can also ease communication and lower barriers. Juan Gilbert's computer and information-science group at the University of Florida in Gainesville, for instance, uses the videoconferencing program Zoom to support lab members during pregnancy and parental leave. Zoom allows them to join lab meetings or have consultations when at home – but only if they choose to. "They want to stay engaged, and using Zoom keeps them connected on their own terms," says Gilbert.

Brenna Hassett, a physical anthropologist at University College London, says such open, transparent group communications can provide a healthy counterbalance to the power dynamics that naturally exist in group meetings led by a principal investigator or a closed-door meeting between a supervisor and their student.

When everyone can weigh in on a conversation, she says, it helps to guard against misunderstandings or misread cues and adds broader context. "You might leave a closed-door meeting thinking your PhD supervisor hates you, when in fact, they just had a reasonable criticism about your bibliography," says Hassett.

Another advantage to all-group communications apps is that they make it harder to say "anything even remotely inappropriate", adds Hassett, who was co-organizer of a session on tech tools for gender inclusion at the 2019 Science Foo Camp conference, held last July in Mountain View, California, and supported by Nature Research, part of Springer Nature (the publisher of *Nature*).

But there are downsides, warns computer

scientist Kate Devlin at King's College London, who co-organized the session with Hassett. "I wonder how many brakes are put on conversations because of the transparency?"

### Research co-op

Another way of flattening group hierarchies is to make research both open and collective.

Marine-data scientist Julia Stewart Lowndes is an advocate of the open-science movement, which espouses open-source software, data sharing and transparency in data analysis and publishing. "But you need a team culture and welcoming environment for people to feel safe" about sharing and discussing data freely, says Stewart Lowndes, who works at the National Center for Ecological Analysis and Synthesis in Santa Barbara, California.

One way to build this trust, she says, is to make codes of conduct or lab values public so that everyone has shared expectations. For instance, one such document from an event that Devlin helped to run states: "We believe that everyone has the right to be in a safe and welcoming environment." The information

**"Any technology that increases communication in a way that is non-threatening, is beneficial."**

might also help to recruit scientists from more diverse backgrounds, if they see that a group welcomes different perspectives and has values that align with their own. (See 'Low-tech tips for inclusivity' online at [go.nature.com/2gdubnt](https://go.nature.com/2gdubnt).)

In Stefania Milan's team, doing science collectively isn't just an attitude, it's an operating principle. Her group at the University of Amsterdam studies the evolution of political activism in the age of big data, and each of the dozen members has an equal say in both group and research decisions.

After receiving a five-year grant to start her group, she and the team spent about 18 months crowdsourcing and developing a set of lab values by which they operate, along with a research questionnaire and the technology infrastructure needed to conduct their research securely. Milan could have completed that work more quickly on her own, she says, but her international and cross-disciplinary team helped to forge a stronger toolkit.

The lab members pool their data and do team analyses using open-source software, including a custom code-sharing system similar to GitHub, which is paired with the ownCloud cloud-storage service. This infrastructure allows them to write and code collaboratively and to share calendars and documents while storing the data on a private, protected server. The whole system is accessible by team members who live abroad or work

from home, so everyone can join in the team's weekly coding session.

### Supporting learning differences

For regenerative pharmacologist Sara Rankin, inclusivity means accommodating a neurodiverse population. Rankin found out late in life that she has dyslexia and dyspraxia: learning differences in the way her brain processes written words and organization. "People work and think in different ways and you've got to allow them to do that," she says.

Rankin's university, Imperial College London, invests in a suite of 16 inclusive software programs to help students and staff who have learning differences or for whom English is a second language. It includes programs such as Grammarly, to check spelling and grammar, as well as tools to help researchers to craft writing or talks in non-conventional ways. Audio Notetaker, for instance, records audio during lectures and syncs it with typed notes, while the speech-to-text software package Dictation.io helps those for whom dictating papers or presentation slides comes more easily.

Rankin uses the idea-mapping program MindView to see her notes on methods together with data charts, images and literature associated with a project – all on one screen. "At a single click, that can be converted into a Word document" as a rough draft of your paper, she says. "That's amazing if you are a visual learner."

Gilbert suggests that leaders take their cues from their team when adopting tech tools such as Slack and WhatsApp. He says that many younger researchers view e-mail as formal and cumbersome. "They are already using these apps in their daily lives, so I wanted to incorporate my lab into that," he says. "And I get a better, more productive workforce that way."

That said, no matter how inclusive an environment might be, there is always room for improvement. Simple technologies can provide anonymous mechanisms for making complaints, reporting inappropriate behaviours or asking questions without the fear of bias or retaliation. For neurobiologist Leslie Vosshall at the Rockefeller University in New York City, an anonymous lab survey proved transformative, she says. The responses prompted Vosshall to make lab meetings more focused and journal club meetings more interactive, and revealed that there was uneven access to lab resources – a problem that was easily solved with a shared online folder for lab protocols.

Embracing that spirit of sharing solutions, resources and power can go a long way towards transforming laboratories into welcoming, fair workplaces. "Science is a social process. We do it in teams and we do it best when we are a diverse, respectful team who care about each other," says Weir. "It's also more fun that way."

**Kendall Powell** is a freelance writer in Boulder, Colorado.





### Where I Work Pamela Yeh

**A**s an evolutionary biologist, my main interest is in understanding how birds and bacteria evolve when they encounter unfamiliar environments. For the past 22 years, I have studied the dark-eyed junco (*Junco hyemalis*), a sparrow with white outer tail feathers. I'm especially interested in a population that has settled at the campus of the University of California, Los Angeles (UCLA). These birds had migrated 70 kilometres from the Angeles National Forest or the woodlands of California's Santa Monica Mountains.

Most of my work is done here on the campus of UCLA, where I am in this picture. I also go with my students into the Californian mountains to see how juncos' appearance and behaviour vary between their natural and their city habitats. There is something so joyful, so wondrous, about going into the on-campus 'field' to study birds – sometimes I feel I know a little secret about the natural world, right here. It makes my heart sing.

Juncos are unusual in that, even in urban settings, they build their nests on the

ground. When I see small shrubs or dense ground cover in some place where people won't tread, I always think, 'Oh, that'd be a good place to put a nest.' If I see a tall tree with branches jutting out, I think that'd be a good place for a junco to sing.

When this photo was taken, my students and I had put up a 'mist net' of fine polyester mesh to capture a junco, and we were trying to lure it by loudly playing recordings of other juncos advertising themselves, attracting mates or defending their territory. In my research, I've found that city juncos are less aggressive and fearful of humans than are mountain juncos. They also have shorter tails and wings, and less white in their tail feathers, and they breed more often.

It's amazing that I can walk from one building to another and see our birds. It's a reminder that nature isn't something we go to – it's where we are. For me, that's very inspiring.

**Pamela Yeh** is an evolutionary biologist at the University of California, Los Angeles.

**Interview by Josie Glausiusz.**

Photographed for *Nature* by  
Sam Comen.

nature

# spotlight

Brain sciences in China

**FOCUS ON  
THE FUTURE**







People at a social-welfare centre interact with the service robot 'A Tie' in Hangzhou, China.

# BRINGING FRESH FOCUS TO AGEING CHINA

Researchers are scrambling to meet the demands posed by a rapidly ageing population. **By Sarah O'Meara**

**W**hen computational biologist Jing-Dong Han moved back to China from the United States in 2005, she found it difficult to convince other scientists there that her research goals were serious.

Han wanted to focus on the science of ageing, an area that did not yet have its own funding stream in China. "Many of my peers told me this was not a real scientific research field," she says.

Han's background in biomedical science, combined with her work as a software engineer

and in computational biology, made her confident that she could prove them wrong. She wanted to deploy the same cellular approaches that she had used in the United States to investigate cancer to probe how the body ages.

"I hoped this information could potentially be used to help people better understand and manage their health needs as they grow older," she says.

China's population is getting old. A decline in the birth rate between 1980 and 2015 as a result of the nation's one-child policy,

combined with increased longevity, will see the proportion of people aged 65 and older triple between 2006 and 2050, to reach almost one-quarter of the population.

The consequences of such a sharp demographic change will be substantial. One-fifth of the population already has neurodegenerative or neuropsychiatric disorders, and the number with age-associated diseases such as stroke, Alzheimer's, Parkinson's and lung cancer will rise (see 'Age-associated diseases'). By 2030, the number of citizens with at least one long-term disorder is expected to be three



China's Yellow Bracelet Project aims to help people with Alzheimer's disease.

times higher than it is today (see [go.nature.com/3dd66bw](http://go.nature.com/3dd66bw)).

For Han, the realization among policymakers in the time since she's moved back to China that the country's health-care system needs to change to cope with the unprecedented demand from elderly citizens has provided momentum for her work. In around 2015, funding agencies in China started to enlist scientists to discuss and propose strategies to address the challenges from population ageing, she says, and that has led to the initiation of dedicated grant tracks and an ever-growing ageing-research community.

Han spent the first five years back in China exploring how genes interact during the ageing process, before being asked in 2010 to lead a pilot project in Shanghai. The Partner Institute for Computational Biology, a collaboration between the Chinese Academy of Sciences (CAS) and the German Max Planck Society, uses computational methods to extract patterns from huge gene and protein data sets to better understand how our cells and body age, among other goals.

In 2015, after publishing numerous studies about ageing in worms and mice, Han and her team published their first work in humans, based on a relatively small cohort of 300 people. It showed how algorithms can use 3D facial images to predict how fast a person is ageing (W. Chen *et al. Cell Res.* 25, 574–587; 2015). She also secured a grant of 1.9 million yuan (US\$270,000) from the Ministry of Science and Technology (MOST). Then, last July, she moved to Peking University in Beijing, and now hopes to publish work on 5,000 individuals.

Ageing research is still very much an

up-and-coming field, says Han, but “even if you can't provide startling results right now or immediately describe how it will be applied, ageing-related research needs to be done urgently in the social context of China”.

### A new field for old problems

Between 2008 and 2018, the number of papers on geriatrics and gerontology published by researchers at Chinese institutions tripled.

In 2016, the Chinese government announced the Healthy China 2030 plan, a national policy to address the country's long-term health-care challenges. The policy includes a range of interventions to promote health in elderly people, such as encouraging a balanced diet and better fitness, and discouraging smoking.

In the same year, the director of the CAS

Institute of Neuroscience, Mu-ming Poo, outlined details of the country's most ambitious neuroscience programme, the China Brain Project, whose remit includes improving our understanding of autism spectrum disorder, depression and neurodegeneration (M.-m. Poo *et al. Neuron* 92, 591–596; 2016).

Then in 2017, the National Natural Science Foundation of China (NSFC) announced a research project with dedicated funding called Organ Aging and Degeneration Mechanisms, through which Han received 2.4 million yuan the following year. And in 2018, MOST launched a programme known as Active Health and Technology Against Aging.

All these efforts are designed to push the development of services and research for senior citizens in China forwards, says Weiping Jia, who runs the Shanghai Institute of Diabetes at Shanghai Jiao Tong University.

The government has also funded centres for basic and clinical ageing research, including the CAS Institute of Stem Cell and Regeneration and the Aging Research Center at Peking University, both in Beijing.

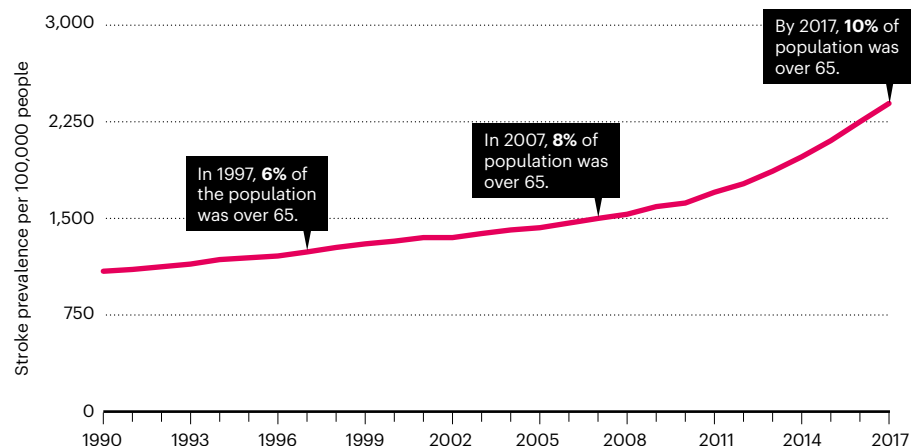
However, James Kirkland, head of the Robert and Arlene Kogod Center on Aging at the Mayo Clinic in Rochester, Minnesota, says that more commitment might be needed, given the size of the problem.

“I think they're on a knife-edge,” he says. “Is this going to be one of the scientific challenges they invest in heavily, or perhaps they'll focus on another research area, such as some form of new space exploration? I don't think they've made that key decision yet to push this kind of systems biology forward as a critical priority.”

Geriatrics scientists in China are keen to tap into the country's massive and homogeneous population by using large-scale studies to generate reliable data on how people age.

### AGE-ASSOCIATED DISEASES

As China's population ages, the country must treat more people with strokes, among other health issues. The proportion of the population aged over 65 is expected to reach **24%** by 2050.



IMAGINECHINA LIMITED/ALAMY

SOURCES: Z. LI ET AL. *BR. MED. J.* 364, 1879 (2019)/WORLD BANK



The nation has already prioritized the use of big data and artificial intelligence in medicine. Its Healthy China plan, for example, will involve building more data centres to collect and combine medical records that are currently spread across separate departments.

### National investment

In 2018, MOST awarded 15 million yuan to Piu Chan, director of the Chinese National Clinical Research Center for Geriatric Disorders at Xuanwu Hospital, Capital Medical University, in Beijing. The grant was so that Chan, a neurologist who researches Parkinson's disease, could use patients' electronic medical records to build a national database of information about neurodegenerative conditions. Given that nearly 60% of people with Parkinson's live in China, that could become a valuable resource for researchers around the world.

Chan is now working with biologist Gang Pei, who is at Tongji University in Shanghai and researches drugs that could reduce the effects of Alzheimer's disease on cognitive decline, to get the database off the ground. It can be tough to convince different government offices to share information, Pei says, because there is little precedent for this kind of national collaboration. "We need to all agree on the rules of how data are shared, how they will be used, what form they come in and who will receive the benefit of any scientific achievement," he says.

Pei would also like to see China develop a repository for tissue samples that can be used by neuroscientists and clinical researchers throughout the nation, as well as worldwide, but says that's likely to be stymied by the stigma attached to organ donation in the country.

"We don't have this tradition. We think the body is sacred. So we need to teach people why they should agree to donate their body to medical science for the greater good."

Guanghui Liu, a stem-cell researcher at the CAS Institute of Stem Cell and Regeneration in Beijing and president of the Chinese Society of Aging Cell Research, says that as well as overcoming cultural challenges, research areas will need to be integrated better if the field is to make progress. "In China, we need cross-disciplinary integration of areas such as biomedicine, physics, chemistry, engineering, bioinformatics and artificial intelligence," he says. "That's probably our greatest challenge to promote the development of the field."

**Sarah O'Meara** is a freelance journalist based in London. Additional research by Kevin Schoenmakers.

## Xiaoming Zhou Neurobiologist

**Xiaoming Zhou is a neurobiologist at East China Normal University in Shanghai. Here he speaks to *Nature* about his research into age-related hearing loss, and explains why he hopes that brain training could help to lessen declines in sensory perception generally, and so ward off neurodegenerative diseases.**

### What is your current research focus?

We want to better understand the neural basis for why a person's hearing function declines as they grow older. For example, we have performed research to see whether we can reverse age-related changes to the auditory systems of rodents.

We gave the animals a set of tasks, such as learning to discriminate between sounds of different frequencies or intensities. These exercises caused the rodents' hearing to improve, and also promoted changes to the hippocampus, a part of the brain structure closely associated with learning and memory.

The relationship with the hippocampus suggests that new kinds of brain training might help to attenuate our declines in perception and other brain functions, such as learning and memory, as we grow older — and so have the potential to stave off neurodegenerative diseases.

### How is ageing-related science developing in China?

As has happened in the rest of the world, a rapidly ageing population has brought significant concern to policymakers. However, as far as I know, only a few neuroscience laboratories in China are specifically focused on learning more about the underlying mechanisms that cause changes in brain function as we age. This is despite the fact that such research could have a considerable impact on the welfare of older people in the future.

Nevertheless, the volume of research carried out in China in this area has increased dramatically — probably because of the huge growth of the country's

scientific community as a whole in recent years, but also because funding for neuroscience research has risen.

### In what areas of ageing-related research is Chinese science making most progress?

I think we could make the most progress in our research on Alzheimer's disease, a neurodegenerative disorder that causes difficulties with memory, cognition and behaviour in many older people. It is one of three brain-related diseases that will be the focus of a forthcoming national neuroscience initiative called the China Brain Project.

So it is foreseeable that a considerable number of Chinese scientists in related fields, such as neuroscience, medicine and artificial intelligence, will work together as part of this plan to study the mechanisms of Alzheimer's. Of course, it's hoped that these studies will receive huge financial support from the central government, and this should help to propel our research.

### What needs to happen for China to make greater leaps in scientific research?

The environment for scientific research, including the management of staff, academic evaluation and financial support, needs to be further improved. For example, researchers spend a lot of valuable time doing administrative work, such as making applications for new instruments and organizing their expenses — time that would be better spent on scientific research, I think.

Fortunately, the management team at my university has noticed this problem and has made a big effort to solve it, and the situation is now gradually improving. In the field of brain science, many Chinese scientists also need more opportunities to collaborate with our international peers to help advance our research. Patience and persistence are very important, too, of course.

### Interview by Sarah O'Meara.

This interview has been edited for length and clarity.

